이학박사 학위논문

Machine learning techniques for decoding and utilizing high throughput RNA sequencing data

RNA 시퀀싱 데이터의 해독과 활용을 위한 기계학습 기법

2019년 8월

서울대학교 대학원

협동과정 생물정보학

김 민 수

이학박사 학위논문

Machine learning techniques for decoding and utilizing high throughput RNA sequencing data

RNA 시퀀싱 데이터의 해독과 활용을 위한 기계학습 기법

2019년 8월

서울대학교 대학원

협동과정 생물정보학

김 민 수

Machine learning techniques for decoding and utilizing high throughput RNA sequencing data

RNA 시퀀싱 데이터의 해독과 활용을 위한 기계학습 기법

지도교수 김 선

이 논문을 이학박사 학위논문으로 제출함

2019 년 5 월

서울대학교 대학원

협동과정 생물정보학

김 민 수

김민수의 이학박사 학위논문을 인준함

2019 년 6 월

위 원 장	황대희
부위원장	김선
위 원	한원식
위 원	이슬
위 원	채희준

Abstract

Machine learning techniques for decoding and utilizing high throughput RNA sequencing data

Minsu Kim
Interdisciplinary Program in Bioinformatics
College of Natural Sciences
Seoul National University

In eukaryotic cells, there are several post-transcriptional modification steps such as RNA editing and alternative splicing, until mRNA molecules are fully matured and translated into proteins. Thus, the transcriptome is a complex mixture of various intermediates that are processed in multiple steps. This complex regulatory structure makes it difficult to fully understand the landscape of transcriptome. My doctoral study consists of three studies that enable RNA-seq to be decoded and utilized in terms of RNA editing, alternative splicing, and gene expression.

RNA editing is a post-transcriptional RNA sequence modification performed by two catalytic enzymes ADAR (A-to-I) and APOBEC (C-to-U). RNA editing is considered an important regulatory system that controls a variety of cellular

functions such as protein activation, alternative splicing, and miRNA targeting. Therefore, detecting RNA editing events in RNA-seq data is important for understanding its biological functions. However, it is known that a significant amount of false-positives occur when detecting RNA editing in RNA-seq. Since it is not possible to experimentally validate all RNA editing residues extracted from RNA-seq, a computational model is needed to filter potential false-positive RNA editing calls. RDDpred, an RNA editing predictor based on machine learning techniques, was developed to filter out false-positive RNA editing calls in RNA-seq. It uses prior knowledge bases to collect training instances directly from the input data, and then trains the random forest (RF) predictors that are specific to the input data. RDDpred was tested using two publicly available datasets of RNA editing studies and has shown good performance.

Another complex problem in RNA-seq decoding is spliceomic intratumor heterogeneity (ie, sITH). Intratumor heterogeneity (ITH) represents the diversity of cell populations that make up the cancer tissue. Recent studies have identified ITH at the transcriptome level and suggested that ITH at gene expression levels is useful for predicting prognosis. Measuring ITH levels at the spliceome level is a natural extension. There is a serious technical challenge in measuring sITH from bulk tumor RNA-seq, such as complex splicing patterns, widespread intron retentions, and short sequencing read lengths. SpliceHetero, an information-theoretic method for measuring the sITH of a tumor, was developed to address the aforementioned technical problems. SpliceHetero was extensively tested in experiments using synthetic data, xenograft tumor data and TCGA pan-cancer data and measured sITH successfully. Also, sITH was shown to be closely related to cancer progression and clonal heterogeneity, along with clinically significant features such as cancer stage, survival outcome, and PAM50 subtype.

The last research topic is to develop a machine learning algorithm that defines patient subspaces specific to particular cancer phenotypes based on gene expression data. Since RNA-seq data is high-dimensional data composed of 20,000 or more genes in general, it is not easy to apply a machine learning algorithm. A network that collects information of experimentally verified interaction of proteins is called a Protein Interaction Network (PIN). Tumor2Vec defines the patient subspace by defining the subnetwork communities that interact with each other by applying the Graph Embedding technique to PIN. Tumor2Vec proposed a clinical model by defining a subspace for patients with different lymph node metastases in early oral cancer and found biologically significant features in the PIN subnetwork unit in the process.

Keywords: RNA-seq, RNA editing, Alternative splicing, Gene expression, Machine learning, Information theory, Graph embedding, Dimension reduction, Autoencoder

Student Number: 2013-23006

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In eukaryotic cells, there are several post-transcriptional modification steps, such as alternative polyadenylation, RNA editing, and alternative splicing before mRNA molecules are fully matured and translated into proteins (Figure 1.1) (Xiang *et al.*, 2018). Thus, the transcriptome is a complex mixture containing various transcriptomic variations that are regulated by different modification systems. This complex regulatory structure makes it difficult to fully understand the landscape of transcriptome.

High throughput RNA sequencing (RNA-seq) is a technology that provides a comprehensive profile of the whole transcriptome by reading vast amounts of RNA fragments (Figure 1.2) (Haas and Zody, 2010). Thus, RNA-seq has been used to elucidate associations between biological phenotypes and transcriptomic variations such as gene expression, RNA editing, and alternative splicing. Each of the three transcriptomic variations has been actively studied and found to be associated with a variety of biological phenotypes.

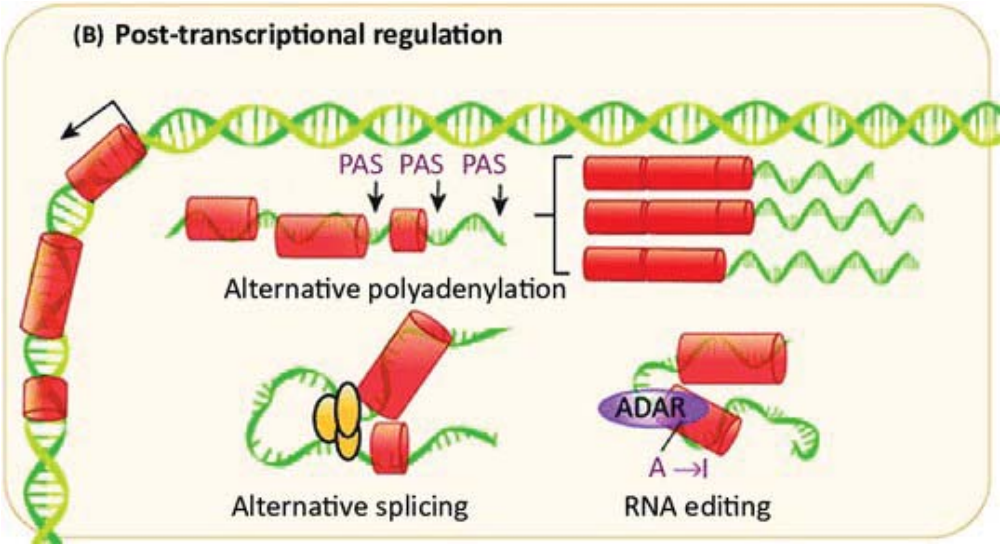

Figure 1.1: A description of the post-transcriptional modification process in eukaryotes (Xiang *et al.*, 2018).

1.1 Biological background

RNA editing

RNA editing is a post-transcriptional RNA sequence modification performed by two catalytic enzymes ADAR (A-to-I) and APOBEC (C-to-U). RNA editing is considered an important regulatory system for controlling various cell functions such as protein activity, alternative splicing, and miRNA targeting (Figure 1.3) (Licht and Jantsch, 2016). There are also several studies showing the direct relationship between RNA editing and biological phenotypes. The study by Galeano et al. suggested that specific RNA editing patterns in glioblastomas by ADAR2 enzymes are crucial for the pathogenesis and that ADAR-class enzymes can be considered as tumor suppressors (Galeano *et al.*, 2013). It is also known that APOBEC3G, a type of APOBEC-class enzyme, causes HIV-1 retroviral inactivation by deamination (Chiu *et al.*, 2010). Therefore, detecting RNA editing

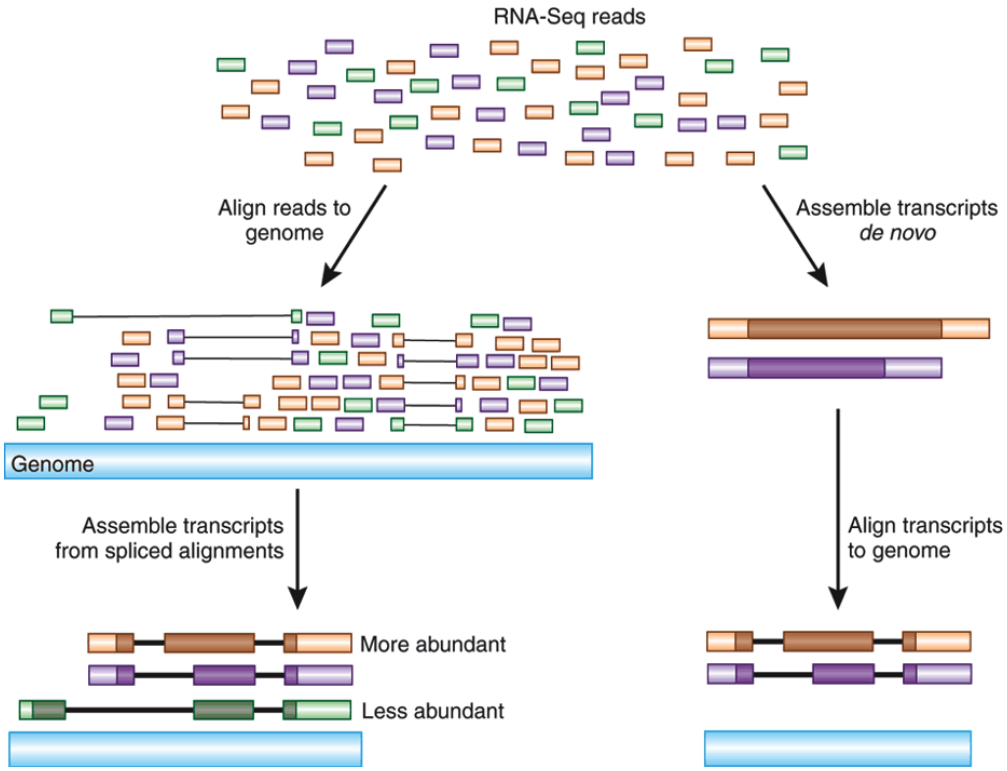

Figure 1.2: A description of RNA-seq data (Haas and Zody, 2010).

events in RNA-seq data is important for understanding the association between RNA editing patterns and biological phenotypes.

Alternative splicing

Alternative splicing is another important post-transcriptional modification that greatly increases the diversity of proteins that can be expressed from a limited number of genes (Liu *et al.*, 2017). The aggregate of cell splicing information is often referred to as spliceome. The term spliceome was coined around 2000 to describe the set of all possible alternatively spliced mRNA and proteins in an organism and all the species depending on the context. Recent studies have

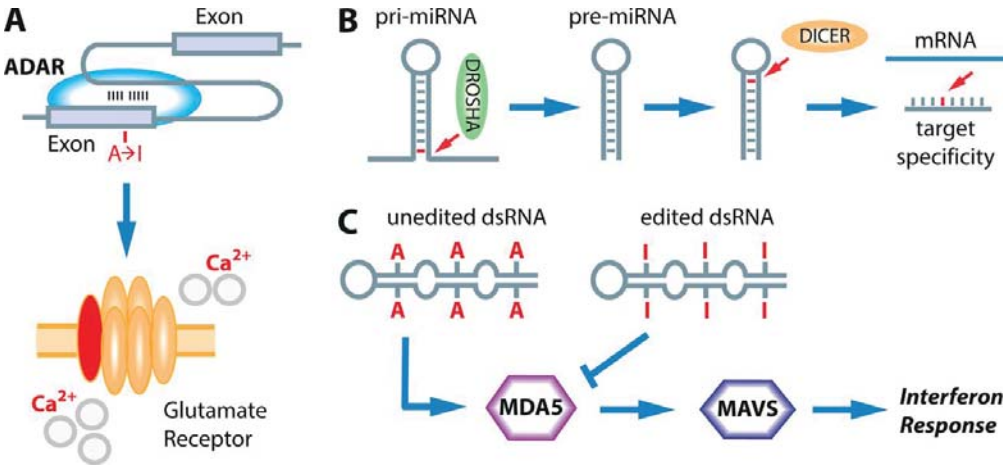

Figure 1.3: A schematic of the intracellular regulatory system governed by RNA editing (Licht and Jantsch, 2016).

suggested associations between cancer phenotypes and spliceomic variations, which can be caused by splice site mutations and malfunctioning splicing factors (Figure 1.4) (Climente-González *et al.*, 2017). RNA-seq is the most effective tool for quantifying spliceome due to its comprehensive profiling capabilities. Many recent spliceome studies have used RNA-seq to understand the biological implications of alternative splicing (Sebestyén *et al.*, 2016; Tsai *et al.*, 2015).

Gene expression

RNA-seq is also a powerful tool that provides gene expression profiles of cells. It uses random primer technology to provide a de novo capture of transcripts without relying on pre-designed probes, which is not possible with microarrays. Because of its high throughput capacity and high resolution, many studies have explored the relationship between biological phenotypes and gene expression patterns using RNA-seq (Figure 1.2) (Haas and Zody, 2010).

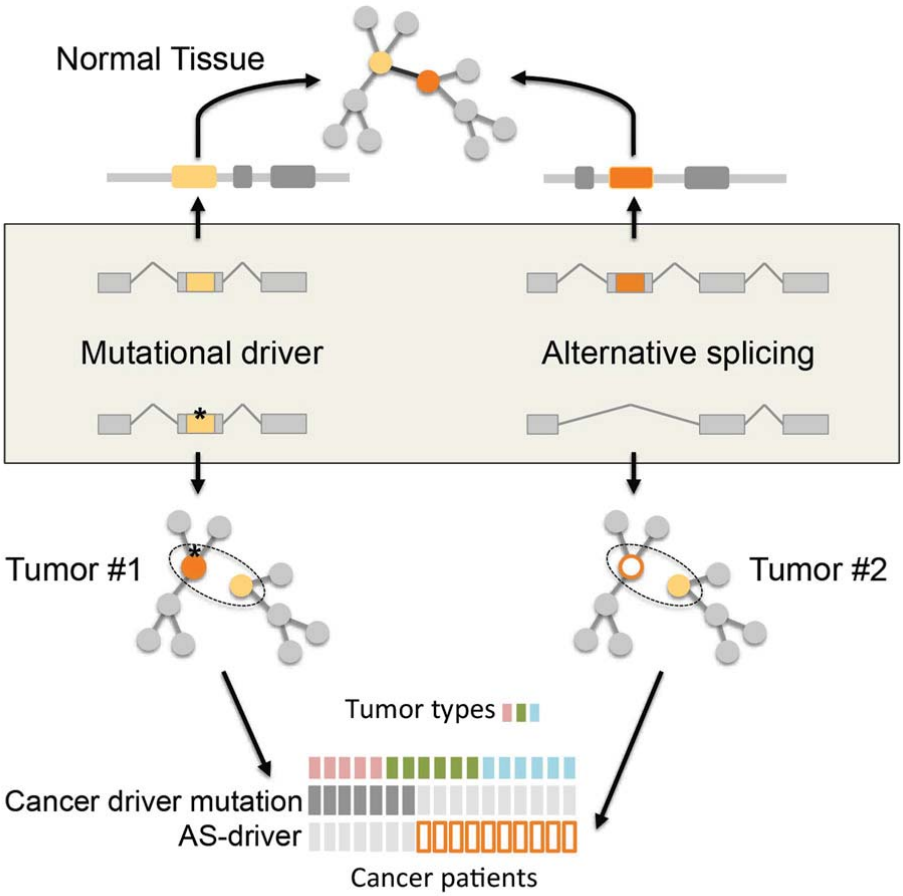

Figure 1.4: A schematic of the intracellular regulatory system governed by RNA editing (Licht and Jantsch, 2016).

1.2 Challenges in decoding and utilizing RNA-seq data

There are challenges in decoding and utilizing RNA-seq data in each of the three transcriptomic domains (ie, gene expression, RNA editing, and alternative splicing). The challenges in each of the transcriptomic domains are summarized as follows.

1.2.1 false-positives in RNA editing calls

It is known that a significant amount of false-positives occurs when detecting RNA editing in RNA-seq. In 2012, Nature Biotechnology published an interview with eight prominent RNA editing researchers in an article called "The difficult calls in RNA editing" (Bass *et al.*, 2012). All eight researchers have pointed out that false-positives are one of the biggest challenges in detecting RNA editing using RNA-seq. They also mentioned that one of the major causes of false-positives is mis-mapping during RNA-seq alignment.

An in silico experiment, part of a preliminary study discussed in Chapter 2, also suggested that mis-mapping poses a significant risk of false-positives (Figure 2.2). Since it is not possible to experimentally validate all RNA editing residues extracted from RNA-seq, a computational model is needed to filter potential false-positive RNA editing calls.

1.2.2 Absence of a model for measuring spliceomic intratumor heterogeneity considering complex cancer spliceome

Intratumor heterogeneity (ITH) represents the diversity of cell populations that make up cancer tissue (Boland and Goel, 2005). This is the result of a subclone diversification process during cancer progression, which is considered a form of Darwinian evolutionary process (Nowell, 1976). The level of ITH reflects the genetic diversity of bulk tumors, which generally have a negative correlation with prognosis. An explanation for this trend is that the genetic diversity provided by ITH can be an accelerator of somatic cell evolution that helps cancer cells acquire a malignant phenotype (Marusyk and Polyak, 2010; Greaves and Maley, 2012; Sun and Yu, 2015; McGranahan and Swanton, 2017).

In a recent study by Morris et al. (Morris *et al.*, 2016), ITH of each cancer sample was first calculated using genomic features such as copy number varia-

Domain	Variation	Method
Genomic	CNVs, Somatic mutations	Mathematical modeling
Methylomic	Methylation	Mathematical modeling
Transcriptomic	Expressional difference	Information theory
Spliceomic	Alternative splicing	None

Table 1.1: Description of each approach using various molecular domains.

tion (CNV) and somatic mutation. Then, the relationship between ITH of each cancer sample and various clinical characteristics was tested. They concluded that the level of ITH in each cancer sample was significantly associated with the molecular, pathologic, and clinical characteristics including prognosis.

ITH can be deduced using molecular profiles of various domains such as genome, epigenome and transcriptome domain. Approaches using each domain have been used to assess the level of ITH in cancer tissues and to identify molecular features associated with tumor evolution (Table 1.1). For example, two ITH studies using genomic variation have revealed somatic mutations that are closely related to tumor evolution in various types of cancer (Carter *et al.*, 2012; Roth *et al.*, 2014). Methylomic and transcriptomic (gene expression) methods for measuring ITH in bulk tumors have been developed and identified important molecular features (Mazor *et al.*, 2016; Park *et al.*, 2016).

The presence of intercellular spliceomic differences has been suggested by studies published over the past decade (Rajan *et al.*, 2009; Wan and Larson, 2018). A recent single-cell study showed that there is a clear difference in the use of isoforms in bone marrow-derived dendritic cells (Shalek *et al.*, 2013). The clinical effect of spliceomic ITH (ie, sITH) has not been thoroughly studied because there is no available sITH model.

There are serious technical challenges in measuring sITH from bulk tumor RNA-seq. A recent study has reported the widespread intron retention of cancer cells (Dvinge and Bradley, 2015), suggesting that the isoform of cancer cells is very complex and not yet characterized. This means that a significant amount of unexpected splice junctions can be found in the cancer sample (Eswaran *et al.*, 2013). To handle these unexpected splice junctions, a transcriptome assembly is required to account for previously unknown isoforms. However, prevalent splice-site mutations (Jayasinghe *et al.*, 2018) and short sequence reads in RNA-seq make it difficult to perform transcriptome assembly. Therefore, a model is needed to avoid this difficulty and to measure sITH in bulk tumor.

1.2.3 Lack of biological interpretation of dimension reduction techniques using gene expression

Transcriptome analysis using RNA-seq is considered to be one of the most effective tools for revealing the underlying biological mechanisms of various cancer phenotypes (Kumari *et al.*, 2017; Lin *et al.*, 2018; Jardim-Perassi *et al.*, 2019). RNA-seq produces a comprehensive expression level of each gene, including more than 20,000, in the case of the human genome. The high resolution of RNA-seq is both an advantage and a cause of trouble at the same time. This problem is also known as high dimension low sample size data problem (McGettigan, 2013; Shen *et al.*, 2016).

There are several machine learning based solutions that address dimensional reduction problems such as Principal Component Analysis (PCA) (Minka, 2001), Latent Dirichlet Allocation (LDA) (Hoffman *et al.*, 2010), Nonnegative Matrix Factorization (NMF) (Cichocki and Phan, 2009), Isomap (Tenenbaum *et al.*, 2000), Locally Linear Embedding (Roweis and Saul, 2000), Multi Dimensional Scaling (MDS) (Kruskal, 1964), Spectral Embedding (Ng *et al.*, 2002), and

t-Stochastic Neighbor Embedding (t-SNE) (Maaten and Hinton, 2008).

Since the existing dimension reduction techniques are unsupervised, the resulting embedding does not reflect the differences between sample labels. Given that the vast majority of cases using RNA-seq are looking for transcriptomic differences between samples with different conditions, these unsupervised approaches do not meet those needs. Also, since the resulting embedding generated by these approaches is generally not provided with a biological interpretation, users must re-process the results in their own way. In this process, the same result is often interpreted differently. Therefore, a supervised learning model is needed that can directly derive the biological interpretation from the resulting embedding.

1.3 Machine learning techniques to solve difficulties in using RNA-seq

- RDDpred, a condition-specific machine learning model for filtering false-positive RNA editing calls in RNA-seq data, was developed to filter out false-positive RNA editing calls in RNA-seq. It uses prior knowledge bases to collect training examples directly from the input data, eliminating the need for expensive experimental verification.

- SpliceHetero, an information-theoretic approach for measuring spliceomic intratumor heterogeneity from bulk tumor RNA-seq data, was developed to solve technical problems caused by complex cancer spliceome. It uses a local analysis approach to avoid transcriptome assemblies that are not easily achievable in cancer spliceome.

- Tumor2Vec, a supervised learning algorithm for extracting subnetwork representations of cancer RNA-seq data using protein interaction net-

works, was developed. It uses the graph embedding technique applied to the PIN to determine the globally well-tuned local subnetwork community. Each community is then considered a feature representation of the input data. It uses machine learning techniques to reduce the dimensionality of RNA-seq data while providing interpretable subnetwork level features.

1.4 Outline of thesis

My doctoral study consists of three studies that enable RNA-seq to be decoded and utilized in terms of RNA editing, alternative splicing, and gene expression. Chapters 2, 3, and 4 introduce independent studies on how to deal with the difficulties of using RNA-seq in each of the three transcriptomic domains.

Chapter 2 describes RDDpred, a condition-specific machine learning model for filtering false-positive RNA editing calls in RNA-seq data, which aims at filtering false-positive RNA editing calls in RNA-seq. Chapter 3 discusses SpliceHetero, an information-theoretic approach for measuring spliceomic intratumor heterogeneity from bulk tumor RNA-seq, which aims to develop a sITH model that takes into account the technical challenges of complex cancer spliceome. Chapter 4 discusses Tumor2Vec, a supervised learning algorithm for extracting subnetwork representations of cancer RNA-seq data using protein interaction networks, which aims to reduce the dimension of RNA-seq data while providing interpretable subnetwork level features.

Chapter 5 summarizes the results of previous studies and the expected results of ongoing studies. This paper is concluded by the bibliography of the references and appendices.

Chapter 2

RDDpred: A condition specific machine learning model for filtering false-positive RNA editing calls in RNA-seq data

2.1 Related works

There are three types of methods for dealing with false-positives in RNA editing calls. 1) Prior knowledge-based filtering, 2) Mapping error prone site estimation, and 3) Machine learning based predictor.

- Prior knowledge-based filtering is the most stringent of all. It collects all potential genomic loci that can cause mis-mapping and excludes all RNA editing residues found nearby (Li *et al.*, 2009; Mo *et al.*, 2014).

- Mapping error prone site estimation is a proactive approach that pre-locates genomic loci that are prone to mapping errors and excludes RNA editing residues from the loci (Peng *et al.*, 2012). To find such loci, they

first synthesize RNA-seq, in which mismatches are intentionally inserted. The generated RNA-seq is mapped to the genome sequence to be evaluated. As a result, mismatches found in areas other than those that were intended at the time of data generation are classified as sites that are prone to mapping errors (Figure 2.1).

- Machine learning-based predictor is a method of using machine learning algorithms to learn the difference between true and false-positive examples and then to determine candidate residues based on the learned model (St Laurent *et al.*, 2013; Zhang and Xiao, 2015).

2.2 Motivation

Compared to the other two approaches, prior knowledge based filtering is relatively naïve and has suspicious performance (Li *et al.*, 2009; Mo *et al.*, 2014). Mapping error prone site estimation has shown better performance (Peng *et al.*, 2012), but it is impossible to simulate all possible conditions. Therefore, it lacks generality. The machine learning-based predictor approach does not suffer from this problem because it can generate generic predictors from training examples (St Laurent *et al.*, 2013; Zhang and Xiao, 2015). Also, according to a study by St. Laurent et al., the predictive accuracy of the machine learning model is quite high (87%) (St Laurent *et al.*, 2013).

One problem with using machine learning approaches in RNA editing calls is that current approaches require experimentally proven RNA editing sites to generate models. Proactively verifying as many sites as necessary for a machine learning model is costly and is not possible if there are not enough samples. Therefore, a machine learning method is required to obtain training examples directly from input data without experimental verification.

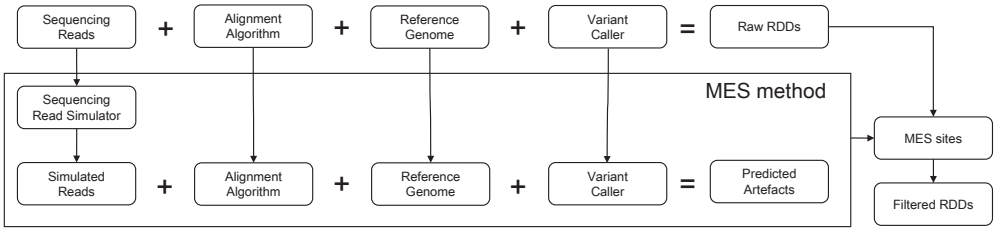

Figure 2.1: A flowchart for the mapping error prone site estimation process (Peng *et al.*, 2012). Mapping error prone sites (MES) are calculated as follows: 1) RNA-seq data with intentionally inserted mismatches are synthesized. 2) The synthesized RNA-seq is mapped to the genome sequence. 3) After SNV call, mismatches found in unintended areas are classified as MES sites.

2.3 A preliminary study

The human genome has lots of loci with mis-mapping risks such as inherent duplicates and repeats, splice sites, and individual polymorphisms. And the short sequence read length of RNA-seq makes all of this worse (Degner *et al.*, 2009; Heap *et al.*, 2010; Engström *et al.*, 2013). A preliminary study was conducted to evaluate the impact of mis-mapping on false-positive RNA editing calls. The test process is as follows (Figure 2.2).

1. Ten RNA-seq data were synthesized, each with 10 million paired-end reads (100bp x 2). Where random single nucleotide variations (SNVs) were inserted with a 1% probability for each read. The inserted 1% SNVs represent individual genetic differences, such as individual SNPs, somatic mutations, and RNA editing.

2. RNA-STAR (Dobin *et al.*, 2013), a state-of-the-art alignment tool in terms of base accuracy (Engström *et al.*, 2013), was used to align the RNA-seq data with the human genome (hg19 build).

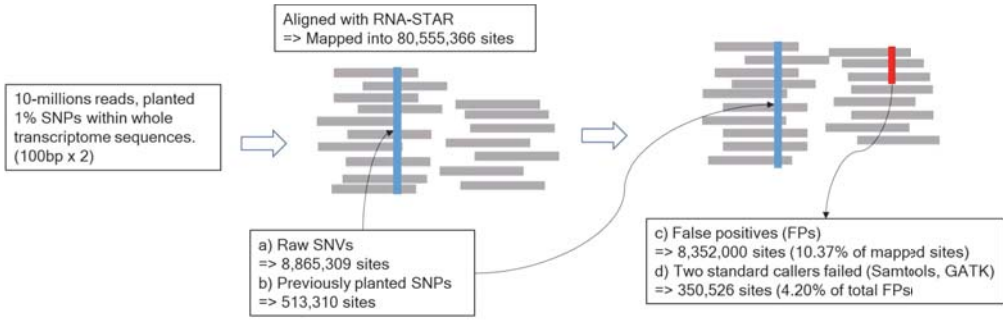

Figure 2.2: A schematic of the preliminary test process.

	Mapped Reads	Mapped Residues	Raw SNVs	false-positives	Standard Caller Failed
ITER_1	9,734,787	80,552,288	8,872,433	8,358,426	350,694
ITER_2	9,735,558	80,558,479	8,878,304	8,365,007	350,670
ITER_3	9,733,473	80,568,898	8,880,681	8,366,553	350,136
ITER_4	9,733,159	80,570,416	8,879,502	8,365,311	350,442
ITER_5	9,733,939	80,545,810	8,853,408	8,339,822	350,332
ITER_6	9,733,507	80,542,007	8,838,870	8,326,074	350,917
ITER_7	9,734,222	80,555,307	8,859,741	8,346,628	350,390
ITER_8	9,735,046	80,562,701	8,874,369	8,361,655	350,807
ITER_9	9,733,971	80,555,609	8,852,720	8,339,866	350,059
ITER_10	9,734,717	80,542,143	8,863,065	8,350,655	350,809
AVG	9,734,238	80,555,366	8,865,309	8,352,000	350,526

Table 2.1: A table for preliminary test results.

3. Standard SNV callers such as samtools (Li, 2011) and GATK (McKenna *et al.*, 2010) have been applied to filter out false-positives in the results.

Table 2.1 summarizes the preliminary test results. On average, 10 million reads result in 8.35 million false-positive residues. Of these, 350,000 residues (4.20%) could not be filtered by overlapping two standard SNV callers (ie, samtools and GATK). The result suggests that when producing 10 million RNA-seq reads to detect RNA editing, there is a risk of 350,000 false-positives on average, which are difficult to filter with standard SNV callers.

2.4 Methods

RDDpred uses prior knowledge bases to extract positive and negative training examples from input data. The whole process is as follows (Figure 2.3).

Input preparation

The RNA-seq data must be properly aligned and converted to BAM file format for input to RDDpred. After receiving the RNA-seq data, RDDpred processes each data using the built-in standard SNV caller samtools-bcftools (Li, 2011). SNVs detected by the standard SNV caller samtools-bcftools are considered candidates for RNA editing.

Preparation of positive training examples

There are two well-organized RNA editing databases called DARNED (Kiran and Baranov, 2010) and RADAR (Ramaswami and Li, 2013). RDDpred queries each RNA editing candidate for each database and considers the residues contained in the database as positive examples.

Preparation of negative training examples

As mentioned, mis-mapping is a major cause of false-positives in RNA editing calls (Bass *et al.*, 2012). Therefore, RDDpred prepares negative training examples using the mapping error prone site estimation method (Figure 2.1) (Peng *et al.*, 2012).

Input feature description

RDDpred constructs a random forest model using 15 features that reflect the local read alignment pattern. The local read alignment pattern is the local state

of the alignment near the SNV. There are at least six categories of attributes calculated from local read alignment patterns such as Read Depth, Allele Segregation, Mapping Quality, Read Position, Base Quality, and Read Strand. The samtools-bcftools pipeline provides 15 statistics in six categories (Table 2.2). Each of the 15 features in the six categories has the following meanings.

- The Read Depth category contains the ReadDepth attribute, which indicates the number of RNA-seq reads that cover each SNV residue.
- The Allele Segregation category contains four attributes, including VAF, SGB, FQ, and CallQual, and is calculated based on the allele ratio of each SNV.
- The Mapping Quality category includes four attributes, PV3, MQB, MQ0F, and MQ, which are calculated based on the quality of the mappings generated by the aligner and generally indicate whether the reads are multi-mapped.
- The Read Position category contains three attributes, including VDB, RPB, and PV4, which indicate how the relative position of the SNV is biased for each RNA-seq read. When overly biased, it usually indicates a mis-mapping.
- The Base Quality category includes two attributes, PV2 and BQB, which indicate whether the low-quality base is highly biased for each SNV, where the base quality is evaluated by the sequencing machine.
- The Read Strand category contains the PV1 attribute, which indicates how biased the strand of the read with the SNV is.

Category	Name	Description
Read Depth	ReadDepth	Read depth
Allele Segregation	VAF	Variant read ratio
Allele Segregation	SGB	Segregation based metric
Allele Segregation	FQ	Phred probability of all samples being the same
Allele Segregation	CallQual	Variant/reference QUALity
Mapping Quality	PV3	Mapping quality bias
Mapping Quality	MQB	Mann-Whitney U test of Mapping Quality Bias
Mapping Quality	MQ0F	Fraction of MQ0 reads
Mapping Quality	MQ	Root-mean-square mapping quality of covering reads
Read Position	VDB	Variant Distance Bias for filtering splice-site artefacts in RNA-seq data
Read Position	RPB	Mann-Whitney U test of Read Position Bias
Read Position	PV4	Tail distance bias
Base Quality	PV2	Base quality bias
Base Quality	BQB	Mann-Whitney U test of Base Quality Bias
Read Strand	PV1	Read strand bias

Table 2.2: A table for the 15 input variables used in the Random Forest model.

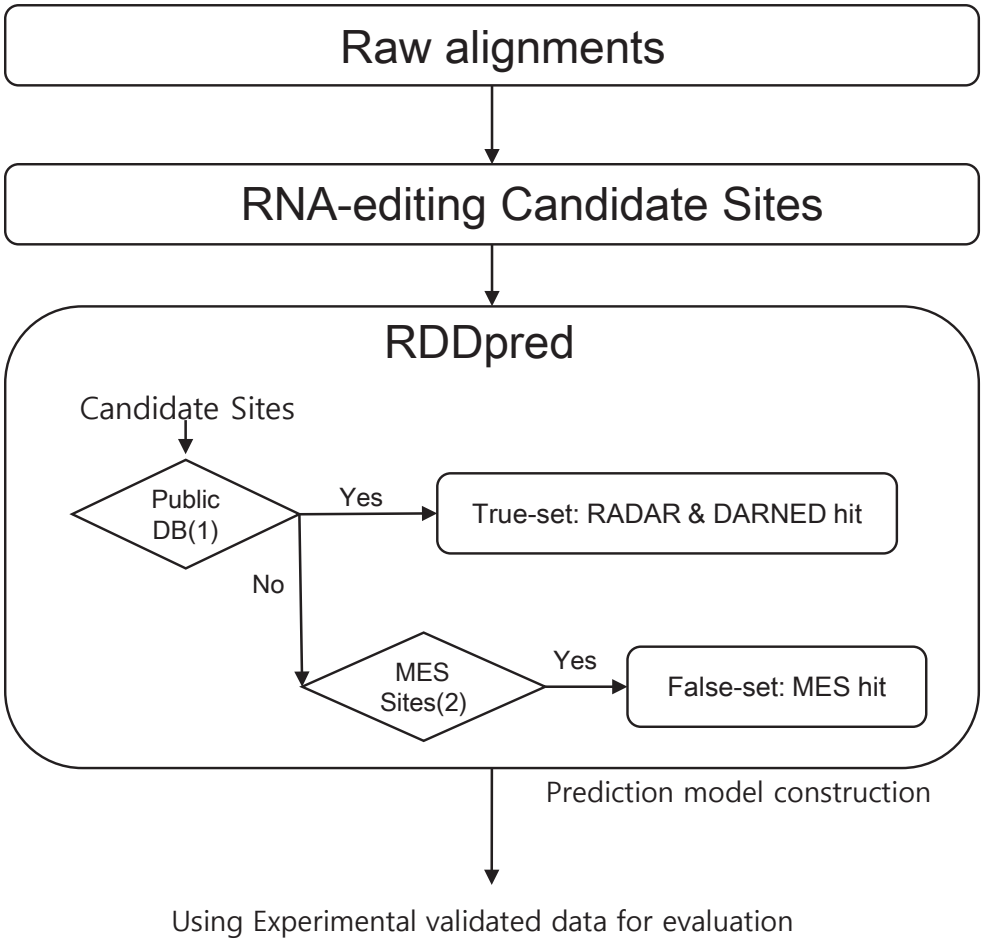

Figure 2.3: A flowchart of total workflow for RDDpred.

2.5 Results

2.5.1 Design of experiments for evaluation

RDDpred was evaluated using the results of two previous studies conducted by Bahn et al. and Peng et al., respectively (Peng *et al.*, 2012; Bahn *et al.*, 2012). Both studies computationally predicted RNA editing sites and validated them with Sanger-seq. The details of the two studies are as follows.

- In the study by Bahn et al. (SRA accession: SRP009659), they collected samples of human glioblastoma astrocytoma and generated 115,132,348 RNA-seq reads. After processing the reads, they predicted 4,141 RNA editing residues as true editing, of which 47 residues were tested with Sanger-seq. They found that 19 residues (40.43%) were false-positives.

- In the study by Peng et al. (SRA accession: SRP007605), they collected samples of human lymphoblastoid and generated 583,640,030 RNA-seq reads. After processing the reads, they predicted 22,688 RNA editing residues as true editing, of which 123 residues were tested with Sanger-seq. They found that 29 residues (23.58%) were false-positives.

2.5.2 Evaluation using data from Bahn et al.

RDDpred detected 6,856,440 RNA-DNA differences (RDD) as a result of primary detection in the 115,132,348 RNA-seq reads produced by Bahn et al. Here, RDD means SNV not found in matched DNA-seq but found only in RNA-seq. RDDpred filtered 6,750,876 residues (98.46%) and predicted the remaining 105,564 residues as true editing.

Overall, the RDDpred results included 3,947 (95.32%) of the 4,141 residues reported by Bahn et al. In the residues tested with Sanger-seq, the result contained 18 of the 47 residues (38.30%), of which 3 were false-positives (16.67%) (Table 2.3).

2.5.3 Evaluation using data from Peng et al.

RDDpred detected 58,666,976 RNA-DNA differences (RDD) as a result of primary detection in the 583,640,030 RNA-seq reads produced by Bahn et al. Here, RDD means SNV not found in matched DNA-seq but found only in RNA-

Bahn et al.	RDDpred Accept	RDDpred Reject	Sum
True Positive	15	13	28
false-positive	3	16	19
Sum	18	29	47

Table 2.3: A table for evaluation results using data of Bahn et al.

Peng et al.	RDDpred Accept	RDDpred Reject	Sum
True Positive	73	21	94
false-positive	7	22	29
Sum	80	43	123

Table 2.4: A table for evaluation results using data of Peng et al.

seq. RDDpred filtered 6,750,876 residues (94.76%) and predicted the remaining 3,076,908 residues as true editing.

Overall, the RDDpred results included 20,504 (90.37%) of the 22,688 residues reported by Peng et al. In the residues tested with Sanger-seq, the result contained 80 of the 123 residues (65.04%), of which 7 were false-positives (8.75%) (Table 2.4).

2.6 Discussion

Evaluation using the results of previous two studies

Overall, the RDDpred results included most of the residues reported in the two studies (Bahn: 95.32%, Peng: 90.37%). This means that RDDpred has successfully reproduced their results. Also, RDDpred results in residues tested with Sanger-seq contained significantly fewer false-positives compared to the residues reported in each study (Bahn: 40.43% $\Rightarrow$ 16.67%, Peng: 23.58% $\Rightarrow$

8.75%). This means that RDDpred has more robust performance than previous approaches. Note that in both comparisons, the residues tested with Sanger-seq were excluded from the training phase of RDDpred for a fair comparison.

Evaluation of feature importance

The 15 input features were evaluated for their ability to distinguish false-positive RNA editing calls. They were evaluated by calculating the information gain using WEKA (Hall *et al.*, 2009). Table 2.5 summarizes the evaluation results. The top five features are contained in two categories, Allele Segregation, and Base Quality. Allele Segregation represents the number of reads that support SNV, and Base Quality represents the quality of sequencing generated by the sequencing machine (Li, 2011). It means that the most important features distinguishing true and false-positive RNA editing are the SNV allele ratio and the base quality evaluated by a sequencing machine.

Software specification

RDDpred was developed as a software package with WEKA, a data mining package, to train a prediction model (Hall *et al.*, 2009). The Random Forest algorithm was chosen because it showed a good performance in the study by St. Laurent et al. (St Laurent *et al.*, 2013). RDDpred was tested in a Linux environment with Python (2.7.3), Samtools-Bcftools (1.2.1), and WEKA (3.6.12). RDDpred can get input from any type of alignment method that provides BAM format output. However, RNA-STAR is recommended for high overall accuracy and high performance (Engström *et al.*, 2013; Dobin *et al.*, 2013). RDDpred is available free of charge at http://biohealth.snu.ac.kr/software/RDDpred/.

Also, to provide information about actual execution time and memory usage, RDDpred tested with the data of Peng et al. (Peng *et al.*, 2012). RDDpred took

Category	Name	Bahn et al.	Peng et al.	Avg. Rank
Allele Segregation	FQ	0.6124	0.3319	2
Base Quality	PV2	0.4746	0.4611	3
Allele Segregation	VAF	0.5526	0.3268	3.5
Allele Segregation	CallQual	0.5737	0.1958	4
Base Quality	BQB	0.425	0.3428	4
Read Depth	ReadDepth	0.4943	0.2515	4.5
Read Position	PV4	0.234	0.1615	7.5
Read Position	RPB	0.2545	0.0712	8.5
Read Position	VDB	0.0988	0.073	9.5
Mapping Quality	MQ0F	0	0.0785	11
Allele Segregation	SGB	0.0932	0.0368	11.5
Read Strand	PV1	0.1584	0.0216	11.5
Mapping Quality	MQ	0	0.0591	12.5
Mapping Quality	MQB	0.0401	0.0367	12.5
Mapping Quality	PV3	0	0.0137	14.5

Table 2.5: A table for input feature evaluation results.

18.33 hours to process 583,640,030 of 101,787,059,720 bases, which is a level that does not hinder the general bioinformatics study. The machine specifications specified in the experiment are as follows.

- Linux version: Linux version 2.6.32-358.el6.x86_64, CentOS release 6.4
- Memory usage: 20GB in maximum
- CPU usage: 20 cores (Intel(R) Xeon(R) CPU E5645 @ 2.40GHz)

2.7 Conclusion

There are limitations to existing methods such as non-machine learning methods lacking generality and machine learning methods requiring extensive proactive experimental validation. RDDpred is a machine learning technique that overcomes these limitations. It uses prior knowledge bases to extract training samples directly from the input data and then generates machine learning predictors specific to the input conditions. This condition-specific nature makes the model generally have good performance. RDDpred was tested using the results of two previous studies and showed good results by significantly reducing the false-positive rate while reproducing most of the residues reported in both studies.

Chapter 3

SpliceHetero: An information-theoretic approach for measuring spliceomic intratumor heterogeneity from bulk tumor RNA-seq data

3.1 Related works

ITH can be deduced using molecular profiles of various domains such as genome, epigenome and transcriptome domain. Approaches using each domain have been used to assess the level of ITH in cancer tissues and to identify molecular features associated with tumor evolution (Table 1.1). For example, two ITH studies using genomic variation have revealed somatic mutations that are closely related to tumor evolution in various types of cancer (Carter *et al.*, 2012; Roth *et al.*, 2014). Methylomic and transcriptomic (gene expression) methods for measuring ITH in bulk tumors have been developed and identified important molecular features (Park *et al.*, 2016; Mazor *et al.*, 2016).

Genome-level ITH has been extensively studied using bulk tumor sequencing data. ABSOLUTE (Carter *et al.*, 2012) is a genomic ITH model that uses somatic cell mutations and CNV profiles of bulk tumors to infer ITH. ABSOLUTE estimated the optimal values of cancer purity and ploidy using a linear programming technique and then estimated the subclonal genome fraction (ie, ITH). A slightly different approach was used in PyClone (Roth *et al.*, 2014). PyClone used the Bayesian model to define the generative relationship between the number of subclones and the observed genomic variation and then used the Bayesian clustering algorithm to select the optimal number of subclones that best fit the observed data.

Recently, an ITH model using a methylation profile was developed. Methylation does not alter the DNA sequence but is linked to genomic DNA. Thus, the DNA methylation pattern has similar characteristics to the genomic variants. For example, both consider both alleles of each locus corresponding to each pair of homologous chromosomes. When bisulfite-seq is used, the methylated base detection process is similar to somatic mutation. Thus, the methylomic ITH (or mITH) model proposed by Mazor et al. used a mathematical modeling approach similar to the genomic profile based model (Mazor *et al.*, 2016)

A transcriptome-level ITH model was recently developed (Park *et al.*, 2016). They used information theory to estimate ITH in bulk tumors. They proposed an interesting idea to consider ITH as the difference in gene expression distribution between normal tissue and bulk tumors. They first used a curated database of molecular pathways, such as the KEGG database (Kanehisa *et al.*, 2016), to construct a template network and construct a probability distribution for each pathway. The divergence between normal tissue and bulk tumor samples is then calculated by the average Jensen-Shannon Divergence (JSD) of each probability distribution for each pathway. This divergence was considered

to be transcriptomic ITH (tITH) for each sample and was found to be related to clonal evolution and prognostic features.

The presence of intercellular spliceomic differences has been suggested by studies published over the past decade (Rajan *et al.*, 2009; Wan and Larson, 2018). A recent single-cell study showed that there is a clear difference in the use of isoforms in bone marrow-derived dendritic cells (Shalek *et al.*, 2013). The clinical effect of spliceomic ITH (ie, sITH) has not been thoroughly studied because there is no available sITH model.

3.2 Motivation

Bulk tumor RNA sequencing

In this study, ITH was measured using bulk tumor RNA-seq data. bulk tumor RNA-seq is a technique that combines bulk sampling with short-read sequencing. A possible alternative for each part is single-cell analysis and single-molecule real-time sequencing (SMRT-seq).

Single-cell analysis has been improved in terms of stability and efficiency and has been used in many biological studies. This technique is very useful for studying ITH because it provides a molecular profile of each cell composed of bulk tumors (Patel *et al.*, 2014). However, due to patient-to-patient heterogeneity, the reproducible cancer model requires extensive study of a large group of patients. Thus, a single-cell approach is not feasible in this case. Another technology, SMRT-seq, is attracting much attention because of its long read length and lack of bias due to cDNA amplification. However, sequencing errors in SMRT-seq are still a problem and production costs are still very high.

Currently, major cancer consortia such as TCGA produce only bulk tumor RNA-seq data. It is therefore difficult to obtain data with adequate clinical

information using the SMRT-seq platform or single-cell platform. Thus, this study focused primarily on bulk tumor RNA-seq.

Difficulty in measuring spliceomic ITH

Since Park et al. (Park *et al.*, 2016) proposed a good model for ITH at the RNA level, spliceomic ITH (ie, sITH) is naturally defined by extending their method. However, there are serious technical difficulties in extending their method to sITH. First, tITH model by Park et al. (Park *et al.*, 2016) requires a template network to create a probability distribution that can not be used in this case. Also, a recent study has reported the widespread intron retention of cancer cells (Dvinge and Bradley, 2015), suggesting that the isoform of cancer cells is very complex and not yet characterized. This problem is more difficult to solve due to the short length of the RNA-seq read.

If you can assemble full-length transcripts from RNA-seq reads, measuring spliceomic ITH will be much easier, even with cancer cells with complex isoform patterns. However, due to the limited length of the RNA-seq read ($<$ 200-bp), it is very difficult to assemble the entire transcript where all possible combinations of splicing loci should be considered. To solve this problem, an empirical method was used to directly combine two distant positions without searching for all possibilities. This method combines two loci likely to result from the same transcript using known gene annotation information (Trapnell *et al.*, 2010). However, extensive splice site mutations in cancer cells produce many noncanonical splice sites that can not be joined because their gene annotation is unknown (Jayasinghe *et al.*, 2018). This noncanonical site, which can not be assigned directly to a specific transcript, can increase the complexity of transcriptome assembly and the possibility of assembly errors.

Therefore, there is a need to develop a new method for solving the above

problems. The following sections describe the definition (method) and performance (results) of our method.

Local analysis approach

As discussed in the previous section, transcript assembly in cancer is very difficult due to complex splicing patterns, noncanonical splice sites, and short-length sequence reads. Therefore, a local analysis approach was devised to avoid transcriptome assembly. In this scheme, all RNA-seq reads that support the splicing event are locally separated and grouped, with each group corresponding to each intron region (Figure 3.1). Spliced aligners such as RNA-STAR (Dobin *et al.*, 2013) align RNA reads with reference genome sequences and output mapped positions on chromosomes.

Because RNA-seq is derived from mature mRNA transcripts, the spliced region remains a gap in the resulting alignment. The aligner collects the spliced gaps and organizes them into splice sites (ie, the ends of the intron). As a result, the aligner lists the position of each splice site on the chromosome observed in a given RNA-seq and the number of supporting reads. The list of splice junctions extracted from the RNA-seq of each bulk tumor is the input data to construct our model (Figure 3.1).

A local unit is then defined, called an *intronic splicing unit* (or *splicing unit*), which is a collection of splicing events for each intron (Figure 3.1). In this scheme, splice junctions sharing a common splice site are grouped into a single unit. As in Figure 3.1, if three splice junctions are sharing a common splice site upstream and three alternate sites downstream, a splicing unit S consisting of A, B, and C can be defined. Where the input variable is defined by the junction count of A, B and C (ie, CNT_S=(5, 3, 2), Equation 3.1). The probability distribution of S can be obtained by dividing the sum of the total

observations (ie, P_S=(0.5, 0.3, 0.2) (Equation 3.1).

$$P_k(i) = \frac{CNT_k(i)}{\sum_{j=1}^{N_k} CNT_k(j)} \tag{3.1}$$

Where $CNT_k(i)$ represents the number of RNA-seq reads that support i-th alternative splice site in k-th splicing unit. $P_k(i)$ is the fraction of RNA-seq reads that support the i-th splice site of the k-th splicing unit. N_k is the total number of alternative sites in k-th splicing unit.

3.3 A preliminary study

An experiment was conducted to test the effectiveness of the local analysis approach. The key question is whether the ITH measured locally (ie, intron-level ITH) is capable of reproducing ITH at the whole-transcript level. The *TP53* gene was chosen because of its well-known implications for cancer progression and its well-characterized isoform structure. *TP53* has 15 isoforms and 12 exons (O'Leary *et al.*, 2015), which are complex enough to be used in experiments. 1,000 RNA-seq samples containing various combinations of 15 isoforms were randomly generated. The RNA-seq simulation was performed using the well-known NGS-seq generator WgSim (CMD: wgsim -e0 -r0 -R0 -X0 -S0 -A1 -d 500 -s 50) (MIT, 2011). The ITH of each sample was measured by the Shannon entropy of the isoform usage profile, as in the study by Graf et al. (Graf and Zavodszky, 2017). Where the value represents the uncertainty or heterogeneity at the spliceome level (Equation 3.2).

$$ITH_{transcript} = -\sum_{i=1}^{N} P(i) log P(i) \tag{3.2}$$

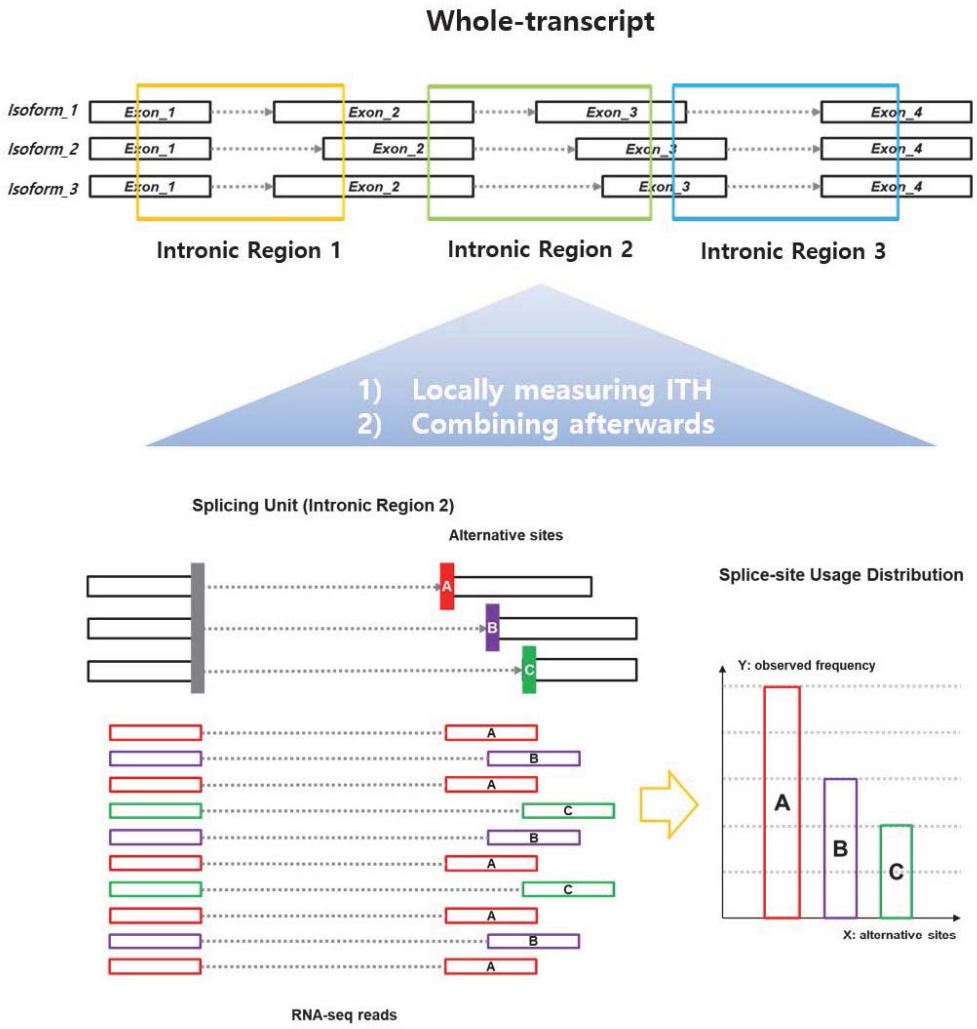

Figure 3.1: An illustration for intronic junction unit. An intronic splicing unit is defined as a set of splicing events that share a common splicing site (ie, donor or receiver) in the intronic domain. Each intronic splicing unit consists of an isoform usage distribution of each sample in each locus. Here, the splice-site usage distribution is calculated by the number of RNA-seq reads that support each alternative splice-site (shown in red, purple, and green in the figure).

Where P represents the *TP53* isoform usage profile for each sample and N is the total number of isoforms. Therefore, $P(i)$ represents the ratio of isoform use of i-th isoform of *TP53*. The $ITH_{transcript}$ represents the whole-transcript level ITH determined by the predefined isoform usage profile, randomly assigned to each sample. On the other hand, the local level ITH was measured using a local splice-site usage distribution (Equation 3.1) extracted from 1,000 RNA-seq samples. ITH_{intron} represents the locally measured ITH defined as Equation 3.3.

$$ITH_{intron} = -\frac{1}{L}\sum_{k=1}^{L}\sum_{i=1}^{N_k} P_k(i) log P_k(i) \tag{3.3}$$

Where P_k represents the isoform usage profile of the k-th splicing unit of *TP53* gene and N_k is the number of isoforms in k-th splicing unit. Therefore, $P_k(i)$ represents the ratio of isoform usage of the i-th isoform in the k-th splicing unit. Finally, L is the number of local splicing units in the *TP53* gene. The locally estimated ITH was shown to successfully reproduce the whole transcript level ITH (Pearson r = 0.66, p = 1.63e-121) (Figure 3.2).

3.4 Methods

Normal tissues are also known to have heterogeneity in the use of isoforms between cells (Shalek *et al.*, 2013). To address this, the spliceomic ITH (ie, sITH) was defined as the distance from the normal tissue sample to the bulk tumor sample. By doing so, the model is expected to eliminate the innate heterogeneity that exists in normal tissues, leaving only the perturbations that occur during cancer progression.

The Jensen-Shannon Divergence (JSD) was chosen to measure the distance between two data points. JSD is defined by averaging bidirectional Kullback-

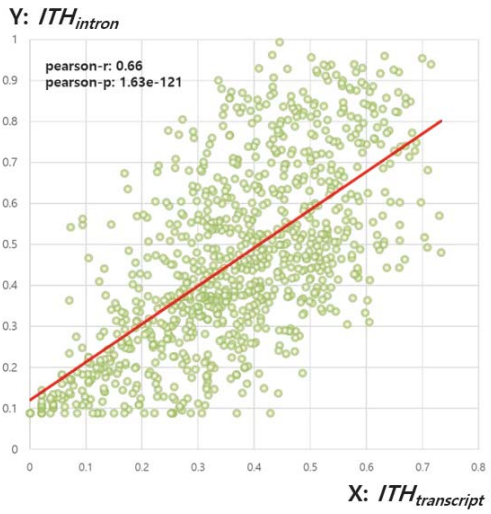

Figure 3.2: A scatter plot showing the correlation between the whole-transcript level ITH and the locally estimated ITH in the *TP53* gene. The X-axis represents the whole-transcript level ITH (ie, $ITH_{transcript}$) and the Y-axis represents the locally estimated ITH by averaging locally measured ITH (ie, ITH_{intron}). Each value is calculated from 1,000 simulated RNA-seq data.

Leibler Divergences (KLDs) from the introduced intermediate data points (Equation 3.4) (Lin, 1991; Joyce, 2011). Then JSD gets the symmetric property and the metric value is limited from 0 to 1 (if you are using a base 2 log). JSD has been used in bioinformatics studies for its symmetric property (Capra and Singh, 2007; Azad and Li, 2012). We have defined input variables representing the distribution of isoform usage for each sample of each locus as a JSD computable form (Equation 3.1).

JSD can be calculated for each intronic region (Equation 3.4). Because each input variable is intended to reflect the use of the splice site at each intronic region, the JSD between the two samples indicates how much the two samples differ in their use of the splice site in that intronic region. This distance is scaled

from 0 to 1. Where 0 means that the splice site usage pattern is the same and 1 is completely different. After calculating the JSD for each splicing unit, a single ITH indicator representing the entire spliceome is calculated by averaging the JSD of all units (Equation 3.7). The detailed calculation procedure of sITH is as follows.

$$JSD(P_k, Q_k) = \frac{1}{2}(KLD(P_k||M_k) + KLD(Q_k||M_k)) \tag{3.4}$$

$$KLD(P_k||M_k) = -\sum_{i=1}^{N_k} P_k(i) log \frac{P_k(i)}{M_k(i)} \tag{3.5}$$

$$M_k(i) = \frac{1}{2}(P_k(i) + Q_k(i)) \tag{3.6}$$

$JSD(P_k, Q_k)$ represents the Jensen-Shannon divergence between the two distributions P_k and Q_k. Where P_k and Q_k denote the splice-site usage distribution of the k-th splicing unit in samples P and Q, respectively. Note that P_k and Q_k are defined in Equation 3.1. M_k represents the intermediate distribution introduced between two distributions P_k and Q_k designed to calculate bidirectional Kullback-Leibler divergence (Equation 3.6). $KLD(P_k \| M_k)$ represents the Kullback-Leibler divergence of the distribution P_k from M_k. Two samples can have different sets of splice sites. In that case, the pseudo-count is added to the splice site, which is not found in one sample, where the pseudo-count is calculated to be 1/100 of the total number of reads in the corresponding splicing unit.

$$sITH(P, Q) = \frac{1}{L} \sum_{k=1}^{L} JSD(P_k, Q_k) \tag{3.7}$$

$sITH(P, Q)$ represents the increased sITH of a sample P from the origin sample Q to be compared. In the actual case, the target sample P corresponds to a bulk tumor sample, and the origin sample Q corresponds to a normal sample. In

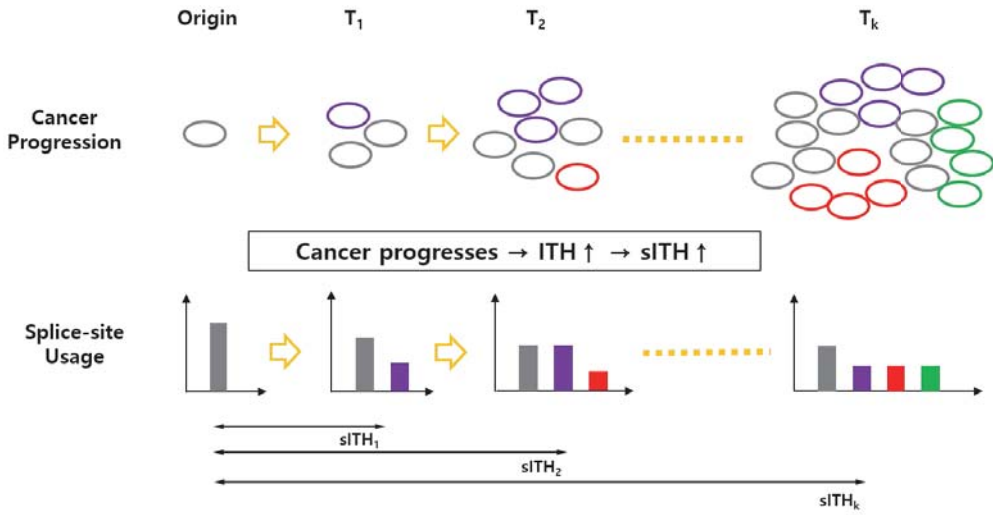

Figure 3.3: An illustration of how cancer progression affects splice-site usage distribution and spliceomic ITH. Clonal heterogeneity increases as a result of cancer progression, which changes the distribution of splice site use in bulk tumors. The sITH is also designed to increase accordingly.

this case, $sITH(P, Q)$ may be called sITH of sample P for convenience (Figure 3.3). L represents the total number of splicing units (usually 20∼30 thousands units found in human cancer tissue). Here, the two samples to be compared are pre-processed using the pseudo counting described above to have the same number of splicing units for compatibility. Therefore, the i-th splicing unit of samples P and Q represents the same intronic region.

The next section is a series of experiments to test whether sITH can function as an ITH indicator and whether it is related to pathological, prognostic, and molecular characteristics.

3.5 Results & Discussion

Three experiments were performed to evaluate the proposed method using 1) synthetic data, 2) xenograft tumor data, and 3) TCGA pan-cancer data.

3.5.1 Synthetic data

The first experiment was performed using synthetic data mixed with normal breast tissue data with single breast cancer data. The purpose of this experiment was to test how the sITH of the mixed sample changes as the mixing ratio increases. The preparation method of the mixture is as follows.

112 Normal breast tissue RNA-seq data were collected from TCGA-BRCA (Network *et al.*, 2012). Then 39 single-cell breast cancer data were collected from a study (SRA accession: SRP159204) (Zhu *et al.*, 2018), where the 39 cells were derived from different clones of a single breast tumor. Each RNA-seq data was processed to obtain the splicing junctions, and the samples were combined in various combinations. Our goal at this stage was to specify a predefined level of ITH in each of the synthetic mixture samples. Initially, normal tissue data were randomly selected from the pool of 112 normal tissues. Then, a certain number of single-cell data were randomly selected, ranging from $1 \sim 39$. The number of selected single-cells represents the ITH level of the mixture. Selected single-cells were mixed into normal tissue at a rate of 1% per cell. For example, if you set the predefined ITH level to 10, the 10 selected single-cell data will be blended into normal tissue data at a 1% rate (10% total) for each cell. Here, the mixing is performed by a weighted sum of the splicing junction counts for each data (Equation 3.8).

$$MIX(i,j) = NT(j) * (1 - i/100) + \sum_{l=1}^{i} (SC(l,j)/100) \tag{3.8}$$

Where i represents the assigned ITH level and j represents the j-th junction of the mixture sample. $MIX(i, j)$ represents the count of the j-th junction of the resulted mixture sample with i ITH level. $NT(j)$ represents the count of j-th junction in the selected normal tissue sample. $SC(l, j)$ represents the number of j-th junction in the l-th selected single-cell cancer sample. Each of the 39 ITH levels was repeated 10 times with random sampling to avoid sampling bias. For example, for ITH level 20, a normal tissue data and 20 single-cell data are randomly selected 10 times each. Thus, a total of 390 mixture samples were synthesized (10 iterations per 39 ITH levels). In conclusion, samples mixed with more single-cells are expected to have a larger ITH by design.

sITH is measured from the origin sample to the target sample distance. In this case, each mixture sample was a target sample, and the normal sample corresponding to each mixture sample was the origin sample. Therefore, the sITH of each mixture shows increased heterogeneity by mixing single-cells. The resulting plot is depicted in Figure 3.4, showing a strong association between the number of mixed single-cells and sITHs (Spearman: r=0.95, p=4.38e-198).

3.5.2 Xenograft tumor data

The main limitation of the previous synthetic data experiment is the lack of an appropriate evolutionary model in the mixture generation. A xenograft tumor data (SRA accession: SRP050242) was used to experiment with conditions that reflect the actual clonal evolution (Chen *et al.*, 2015). The xenograft mouse model used in the experiments originally originated from the human breast cancer cell line (MCF10A).

Cell lines were treated with *HRAS* transduction before transplantation to enhance malignancy. After the single-cell origin derived from MCF10A-*HRAS* was transplanted into immunocompromised mice, the xenograft tissues

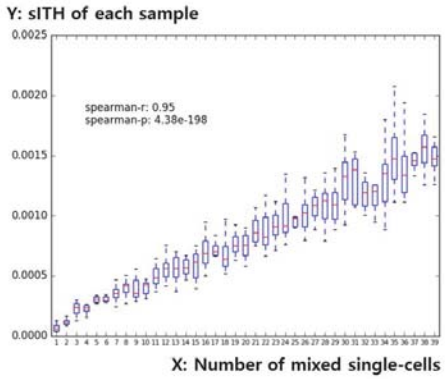

Figure 3.4: A Boxplot to show the association between the number of synthesized single-cells and the sITH of synthesized data. The X-axis represents the number of mixed single-cells (1∼39). The Y-axis represents the sITH of the sample mixed with the number of single-cells specified on the X-axis.

were cultured until the tumor had completely progressed and metastasized. DNA and RNA samples were collected at various points during the process. Thus, the trends in ITH values measured by two different data types (genome and spliceome) can be compared as the tumor grows. A total of 10 samples were collected while culturing xenograft tissue. They collected two samples for metastatic tissue and one sample for each of the eight time-points. For each sample, sITH was calculated from the normal breast tissue samples provided by TCGA-BRCA (Network *et al.*, 2012). Ten randomly selected samples are assigned to each xenograft sample to avoid sampling bias. The sITH of each xenograft sample is then iteratively calculated for each of the 10 normal samples.

The goal at this stage was to test how the sITH of a tumor changes as cancer progresses and the clonal substructure expands. Initially, it was tested how sITH changes over time. As shown in Figure 3.5-(a), sITH has a positive correlation

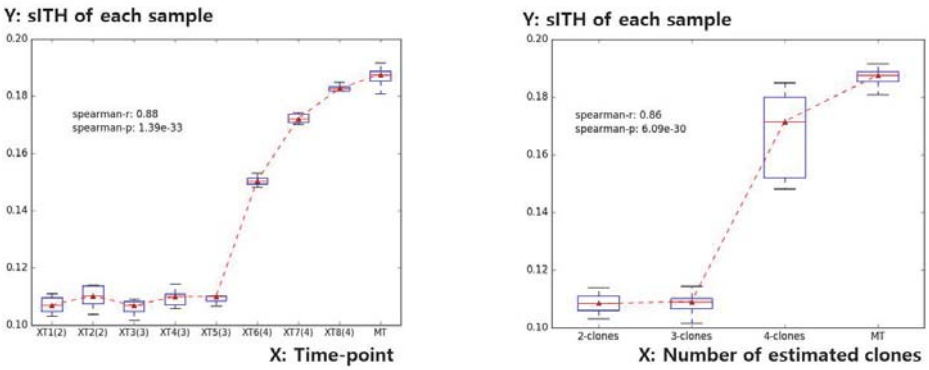

Figure 3.5: Two boxplots of how the xenograft time-point and estimated subclone numbers are associated with sITH. a) The X-axis represents when each xenograft tumor sample was collected. The Y-axis represents the sITH for each sample (including repeated measurements for 10 normal tissues randomly selected for each tumor sample). b) Same as a) except that the X-axis represents the number of subclones estimated by PyClone.

with the time point (Spearman: r=0.88, p=1.39e-33), which means that sITH increases as the cancer progresses. Next, it was tested how sITH changes as the number of subclones increases. The study by Chen et al. (Chen *et al.*, 2015) used PyClone to give the estimated number of subclones in each sample, and these values were compared to sITH. As shown in Figure 3.5-(b), sITH is strongly correlated with the number of subclones (Spearman: r=0.86, p=6.09e-30), which means that sITH increases as the clonal substructure expands.

3.5.3 TCGA pan-cancer data

There are more problems to consider in clinical cases. For example, unlike xenograft samples that share a common ancestral cell, samples from actual cancer patients originated from diverse populations. This heterogeneity between patients due to various genetic backgrounds can be a confounding factor that

might mask the actual ITH. It is also unclear whether the tissue of origin could affect the outcome because only breast cancer tissues were tested in previous experiments. Thus, a comprehensive pan-cancer level experiment was needed to test whether sITH could overcome potential problems and demonstrate clinical significance.

For this purpose, the TCGA pan-cancer dataset was used (Weinstein *et al.*, 2013). The TCGA pan-cancer dataset is a cancer cohort collective that includes 28 cohorts and 9,274 bulk tumor RNA-seq samples (Table 3.1). Corresponding normal tissues are needed to calculate the sITH of bulk tumors, 8 of the 28 groups are excluded because there is no normal tissue. This means that 984 samples were excluded. Overall there are 8,290 available primary tumor samples of 20 types of cancer (Table 3.1). The sITH of each bulk tumor was calculated by processing 8,290 RNA-seq data before experimenting. The sITH of bulk tumor samples in each cohort is calculated using the corresponding normal tissue samples. For example, the TCGA breast cancer cohort (BRCA) has 1,093 primary bulk tumor RNA-seq samples and 112 normal tissue RNA-seq samples. In this case, the sITH of each bulk tumor sample was calculated by averaging the calculated sITHs for each of the 112 normal tissue samples.

The measured sITH values of each bulk tumor sample were compared with clinical features such as genomic ITH (gITH), cancer stage, survival outcome and PAM50 subtype. The following sections describe the comparison procedure and the results.

Comparison with gITH

The gITH used in the experiment is the result of ABSOLUTE (Carter *et al.*, 2012). A study by TCGA provides gITH values for pan-cancer dataset (Weinstein *et al.*, 2013). Of the 8,290 samples with sITH values, 7,594 samples had

DISEASE	NT	PT	sITH	gITH	STAGE	SURVIVAL	PAM50
BRCA	112	1,093	1,086	1,013	995	322	480
KIPAN	129	889	889	659	632	287	0
GBMLGG	5	669	669	644	0	275	0
STES	46	599	599	558	522	232	0
HNSC	44	520	520	485	422	238	0
LUAD	59	515	515	489	487	210	0
LUSC	51	501	501	465	464	245	0
THCA	59	501	501	446	444	97	0
PRAD	52	497	497	469	0	87	0
BLCA	19	408	408	398	396	211	0
COADREAD	51	379	379	351	338	109	0
LIHC	50	371	371	354	333	155	0
CESC	3	304	304	291	0	99	0
SARC	2	259	259	242	0	125	0
PCPG	3	179	179	160	0	31	0
PAAD	4	178	178	158	156	91	0
UCEC	24	176	176	170	0	45	0
THYM	2	120	120	103	0	36	0
SKCM	1	103	103	103	99	29	0
CHOL	9	36	36	36	36	19	0
OV	0	303	0	0	0	0	0
LAML	0	173	0	0	0	0	0
TGCT	0	150	0	0	0	0	0
MESO	0	87	0	0	0	0	0
UVM	0	80	0	0	0	0	0
ACC	0	79	0	0	0	0	0
UCS	0	57	0	0	0	0	0
DLBC	0	48	0	0	0	0	0
SUM	**725(20)**	**9,274(28)**	**8,290(20)**	**7,594(20)**	**5,324(13)**	**2,943(20)**	**480(1)**

Table 3.1: A table for input feature evaluation results.

matched gITH values, and the remaining 696 samples were not provided with gITH values and were excluded from the comparison (Table 3.1).

The Spearman correlation test showed a strong correlation between sITH and gITH (r=0.24, p=6.92e-100). To summarize the vast quantities of results, a percentile boxplot was prepared (Figure 3.6). Where each box contains 10% of the sample in ascending order of gITH. For example, the first box in Figure 3.6 contains samples with gITH rank between 0 and 10%, and the second box contains 10% to 20% samples. The binned representation of Figure 3.6 is used only for visualization, and the actual correlation test is performed by directly comparing the sITH and gITH values of each sample.

gITH is the current golden standard for ITH levels in bulk tumors. Thus, it was used as a reference standard for sITH in all of the following comparisons

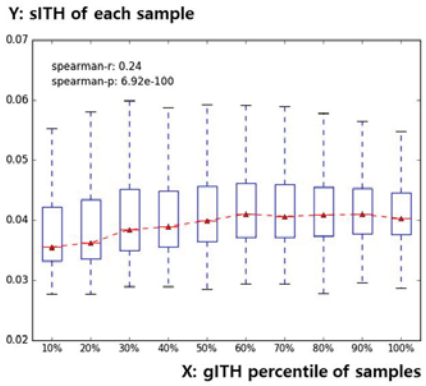

Figure 3.6: A boxplot representing the relationship between gITH and sITH. The X-axis consists of 10 bins that evenly divide the entire sample. Each bin corresponds to a 10 percent scale percentile, ordered by the gITH value of each sample. The result indicates a significant correlation between gITH and sITH.

Comparison with cancer stage

The cancer stage is a well-known indicator of cancer progression, which is determined based on pathological observations of cancer tissues such as size, location, the extent of invasion, and extent of spread. The level of ITH is generally related to the progression of cancer.

Of the 7,594 samples containing both sITH and gITH, 5,324 samples also have cancer stage information (Table 3.1). Figure 3.7 summarizes the correlation between sITH, gITH and cancer stage in each sample. Both ITHs showed a significant correlation with cancer stage, but sITH showed better association (gITH: r=0.11, p=8.11e-17, sITH: r=0.24, p=2.43e-68). The result means that the samples with higher cancer stages have a larger sITH value.

Association with survival outcome

Overall survival represents the survival time after treatment, which is the surgical resection of the tumor in this context. The level of ITH in the tumor is associated with the degree of malignancy of cancer, which in turn affects the mortality rate of cancer patients (Morris *et al.*, 2016). Therefore, the relationship between sITH, gITH and the survival outcome of each sample was tested.

The Cox proportional hazards (Coxph) model was prepared to test 7,594 samples with both sITH and gITH (table 3.1). The Cox regression model is designed to quantify the effect of sITH and gITH on overall survival, where the magnitude of the association is expressed as a p-value. The "CoxPHFitter" function used in the experiment is included in the Python library lifelines (0.21.0). The results of the analysis are summarized in Table 3.2.The results show that sITH is significantly associated with overall survival (HR=1.85e+23, p =1.04e-64). It is better than gITH (HR=3.8, p=1.95e-28).

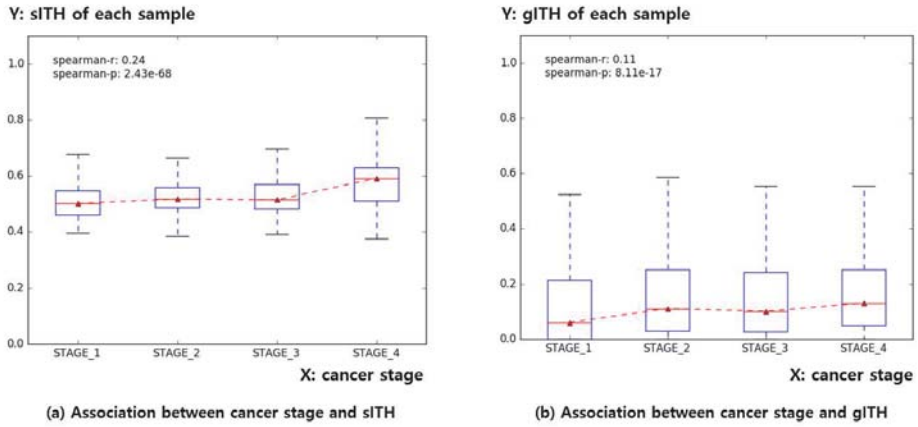

Figure 3.7: A boxplot showing the association of sITH, gITH and cancer stages in each sample. a) The X-axis represents the cancer stage of each sample (1 to 4 stages). The y-axis represents the sITH value of each sample. b) Same as a), but in this case, the Y-axis represents the gITH value of each sample. The results show that both sITH and gITH have a significant correlation with the cancer stage, and the significance is greater in sITH. The sITH and gITH values were standardized by dividing the maximum value between samples so that the distribution of the data is easily understood.

An additional analysis was prepared to help visual understanding. Initially, 7,594 samples were classified into six groups with different survival outcomes. The first five groups were classified by the time of death. For example, the first group contains samples that died in the first year after treatment, and the second group contains samples that died in the second year. The sixth group includes samples reported to be alive for more than 5 years, where the 5-year threshold is based on criteria commonly used to determine cancer remission. As a result, 2,943 samples were classified into six survival groups and the remainder were excluded because they could not be classified into six groups because of the short follow-up period (Table 3.1).

	coef	exp(coef)	se(coef)	z	p_value	lower_0.95	upper_0.95
sITH	53.57	1.85E+23	3.15	16.99	1.04E-64	47.39	59.76
gITH	1.34	3.80	0.12	11.06	1.95E-28	1.1	1.57

Table 3.2: A table for Cox proportional hazards analysis results.

Figure 3.8 summarizes the association between sITH, gITH and the survival group of each sample. Both gITH and sITH were significantly correlated with survival groups, whereas sITH showed better association (gITH: r=-0.20, p=1.75e-27, sITH: r=-0.27, p=6.44e-51). The result indicates that sample groups having higher lethality have a tendency to have greater sITH. The sample information for each sample group is summarized in Table 3.1.

Association with PAM50 subtype

One of the most studied cancer types in terms of the molecular level is breast cancer, and breast cancer has a well-known molecular subtyping system, PAM50 ((Parker *et al.*, 2009)). PAM50 classifies breast tumors into four types: Luminal A, Luminal B, Her2-enriched, and Basal. The order here indicates the degree of malignancy. The associations of sITH, gITH, and PAM50 subtypes were tested.

Of the 1,013 breast cancer samples available for both sITH and gITH, 480 samples have PAM50 subtype information. Both ITHs showed a significant correlation with the PAM50 subtype (Figure 3.9), while sITH showed a better correlation (gITH: r=0.36, p=2.91e-16, sITH: r=0.61 , p= 1.11e-48). The experiment results indicate that groups of samples that are expected to be more malignant by molecular subtypes tend to have a higher sITH.

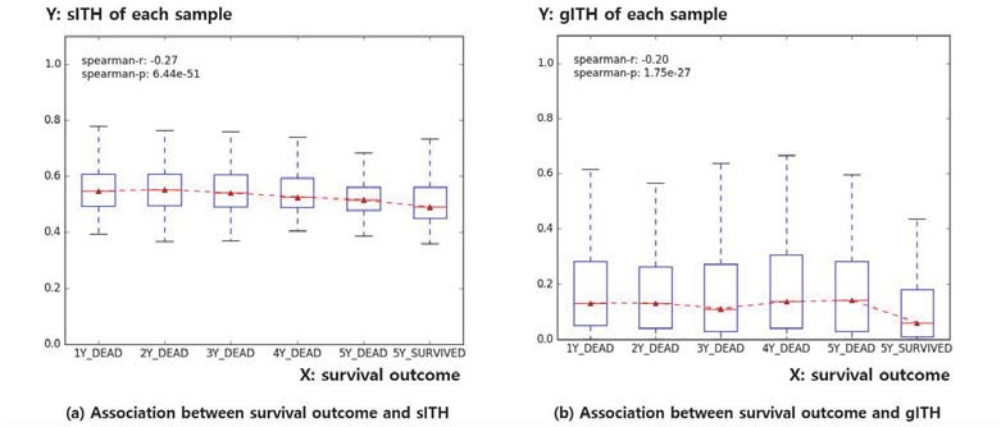

Figure 3.8: A boxplot indicating the association between sITH, gITH and the survival outcome of each sample. a) The X-axis represents a sample population that is classified into the overall survival results of each sample ($1Y_DEAD \sim 5Y\_DEAD$, and 5Y_SURVIVAL). For example, the 1Y_DEAD group represents a sample that dies within one year of surgery. Similarly, 2Y_DEAD corresponds to samples that died within two years of treatment. The remaining groups are defined accordingly. Finally, the 5Y_SURVIVAL group represents the samples that survived 5 years after surgery. The Y-axis represents the sITH value of each sample. b) Same as a). However, this time, the Y-axis represents the gITH value of each sample. Both sITH and gITH showed a significant correlation with survival, but sITH showed a better correlation than gITH. The sITH and gITH values were standardized by dividing the maximum value between samples so that the distribution of the data was easily understood.

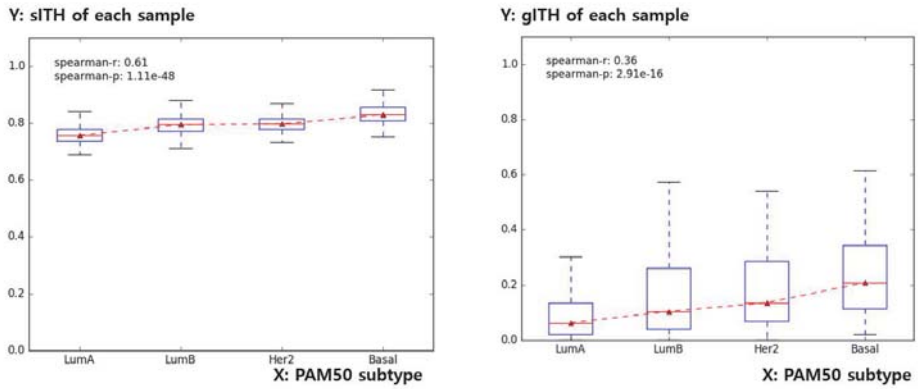

Figure 3.9: A boxplot indicating the association between sITH, gITH and the PAM50 subtype of each breast cancer sample. a) The X-axis represents the PAM50 subtype of each sample sorted by the known malignancy order of each subtype. The Y-axis represents the sITH value of each sample. b) Same as a). However, this time, the Y-axis represents the gITH value of each sample. Both sITH and gITH showed a significant correlation with PAM50 subtype, but sITH showed a better correlation than gITH. The sITH and gITH values were standardized by dividing the maximum value between samples so that the distribution of the data was easily understood.

3.6 Conclusion

Despite studies that show intercellular differences at the spliceome level(Shalek *et al.*, 2013; Wan and Larson, 2018), the clinical effect of sITH has not been studied sufficiently because there is no sITH model. SpliceHetero is a sITH model based on local analysis approach that avoids transcriptome assembly which is not easy in cancer RNA-seq. The proposed model was extensively tested for its performance using synthetic data, xenograft tumor data, and TCGA pan-cancer data. As a result, sITH has shown a strong association with

cancer progression and clonal heterogeneity as well as clinically relevant features such as cancer progression, survival outcome, and PAM50 subtype. Also, the distribution of sITH values within each sample group appears more strict than gITH (Figure 3.7, Figure 3.8 and Figure 3.9). That means sITH is a more consistent indicator than gITH.

The proposed model can help to develop diagnostic and prognostic tools by providing a tool to understand the inherent heterogeneity of cancerous spliceome. The whole process is implemented as a software package and is available free at http://biohealth.snu.ac.kr/software/SpliceHetero. It was implemented in Python 2.7 and tested on CentOS Linux release 7 and Ubuntu 16.04, and 18.04.

Chapter 4

Tumor2Vec: A supervised learning algorithm for extracting subnetwork representations of cancer RNA-seq data using protein interaction networks

4.1 Related works

Precision cancer medicine is a new form of medical practice that provides optimal treatment for each cancer patient by considering the genetic and molecular background as well as clinical history and pathology (Figure 4.1). The basic idea is that medical decisions for each patient can be made by considering the treatment records of previous patients with similar molecular profiles. Thus, one of the major challenges of precision cancer medicine is defining a patient subspace, where the patient-patient distance is defined based on the molecular profile.

RNA-seq is one of the most promising techniques for extracting whole-

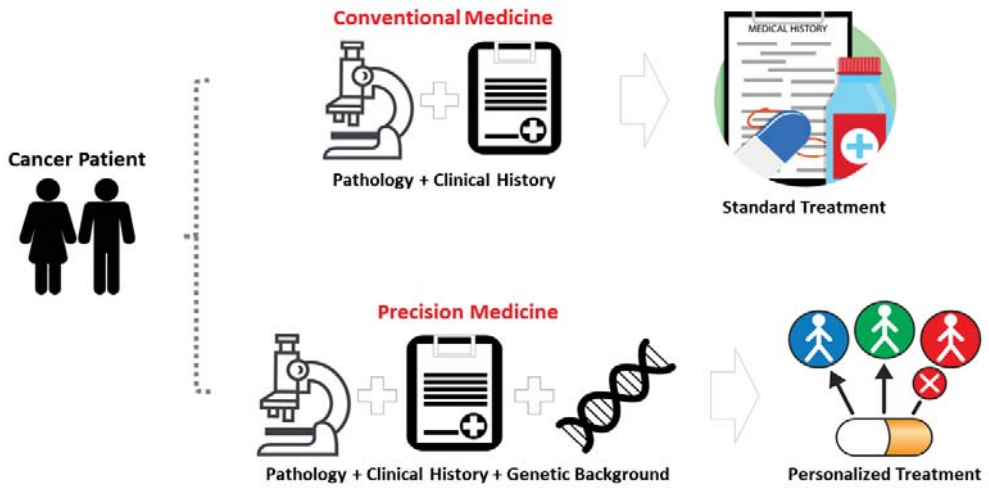

Figure 4.1: An illustration for describing precision cancer medicine.

transcriptome profiles of cancer patients. However, the high-dimensional nature of RNA-seq data (more than 20,000 genes to consider) makes it difficult to define the optimal feature representation that can characterize each patient (Figure 4.2) (McGettigan, 2013; Shen *et al.*, 2016). Because the cost of producing RNA-seq data is still significant, a solution is needed to reduce the dimensionality of the data. There are two main approaches to dealing with this problem.

1. The unsupervised dimension reduction approach mathematically eliminates data redundancy and provides component values that represent each reduced embedding dimension as feature values.

2. The network-based transcriptome analysis approach removes the redundancy of data by grouping genes into subnetwork modules using protein interaction networks and then integrating biologically interdependent genes into a single feature.

The unsupervised dimension reduction approach has been used to reduce

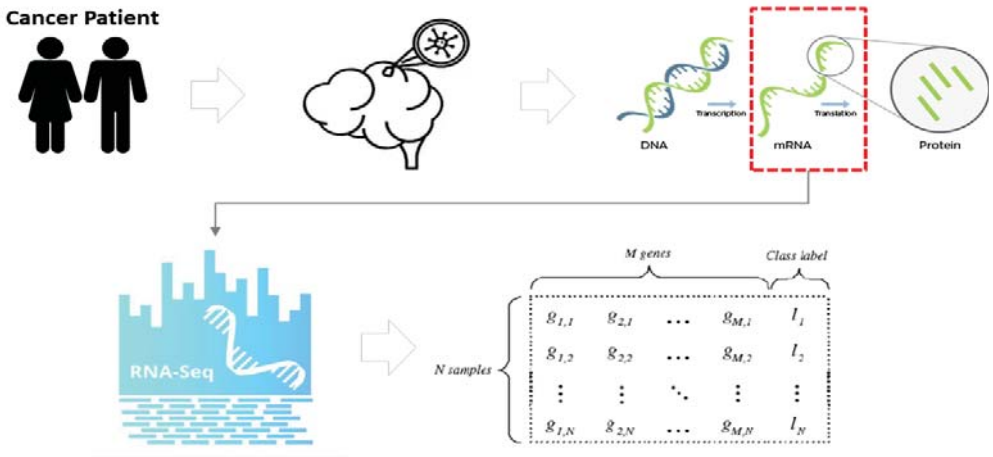

Figure 4.2: An illustration for describing the high-dimensionality issue in RNA-seq.

the data dimension of RNA-seq and has shown particularly good performance when visualizing a collection of data (Treutlein *et al.*, 2014; Wang and Gu, 2018). One limitation of this approach is that it does not provide a biological interpretation of the results. Researchers have to make their own interpretations, and sometimes the same data can lead to different conclusions depending on the interpreter. The network-based transcriptome analysis approach has the advantage that it provides intuitively interpretable subnetwork level features.

Subnetworks defined by protein interaction networks have been associated with various biological phenotypes using a systems biology approach. Most current approaches, however, rely heavily on feature engineering, which requires domain expertise and manual curation (Yu *et al.*, 2013; Xiong *et al.*, 2017; Fan *et al.*, 2018). Two approaches have recently been introduced to extract subnetwork features associated with specific biological phenotypes in an automated manner.

1. Yuan et al. (Yuan *et al.*, 2017) introduced a greedy search algorithm

to find subnetwork features in a protein interaction network (PIN). They defined subnetwork features starting from one gene to finding locally maximized boundaries in terms of defined perturbation scores.

2. Lin et al. (Lin *et al.*, 2017) used a neural network model to find network-based feature representations. They used knowledge bases containing gene regulation structures such as TF networks and PINs to build selectively connected neural network architectures. The internal weights, which are calculated as a result after training the neural network, are considered to be network-level features.

4.2 Motivation

The approach of Yuan et al. (Yuan *et al.*, 2017) considered each local subnetwork as an independent variable in assessing the impact on the corresponding biological phenotype and did not consider their interactions. Therefore, this approach has limitations in dealing with complex diseases such as cancer, because, in cancer, two or more intracellular processes interact to produce a cancer phenotype (Prahallad and Bernards, 2016). The approach of Lin et al. (Lin *et al.*, 2017) is free from this problem because it considers interactions between local subnetworks by using a fully connected neural network architecture. However, despite its high performance, this model has the limitation that it is not easy to interpret its data representation. Therefore, there is a need for a supervised learning model that automatically extracts features of the subnetwork level with biological interpretability.

4.3 Methods

Extraction of local subnetworks

Tumor2Vec uses the graph embedding technique applied to the PIN to determine the globally well-tuned local subnetwork community. Each community is then considered a feature representation of the input data. The process is as follows.

- First, protein interaction information is extracted from a well-organized PIN database STRING. (Szklarczyk *et al.*, 2018). The PIN graph is then constructed from that information.
- The PIN graph is then processed by the graph embedding algorithm DeepWalk (Figure 4.3) (Perozzi *et al.*, 2014). Here, the graph embedding algorithm is performed to find globally well-tuned local subnetwork communities. Through graph embedding, each gene is transformed into an embedding space where the intergenic distance represents the random walk probability distribution of the original PIN graph (Perozzi *et al.*, 2014).
- K-means clustering is applied to all genes to find clusters, where the intergenic distance is measured using the coordinates in the embedding space. Thus, the resulting clusters represent the local connection of the original PIN graph (Perozzi *et al.*, 2014). These clusters are considered local subnetwork features (Figure 4.4).

Training of the kernel function

Recent supervised learning methods that rely on network-based features typically use explicit models (Conte *et al.*, 2013) that learn functions that map input

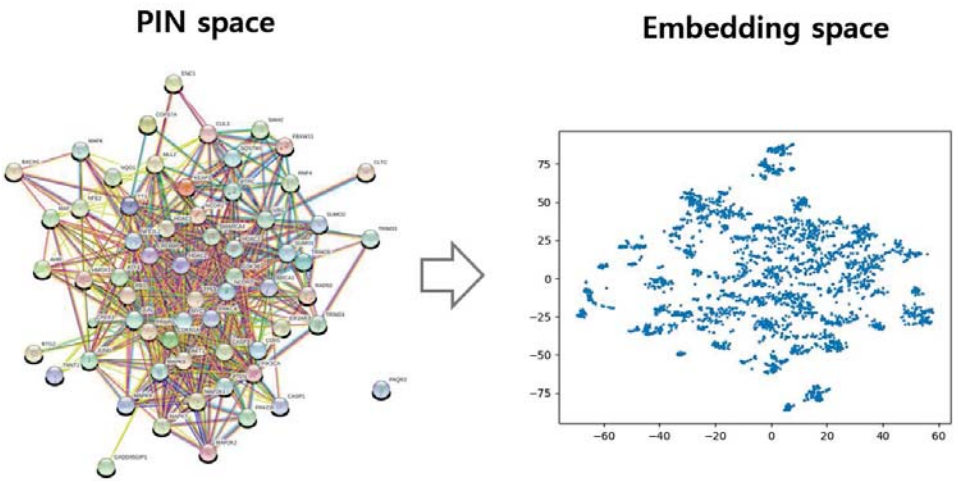

Figure 4.3: An illustration of the graph embedding process (Perozzi *et al.*, 2014).

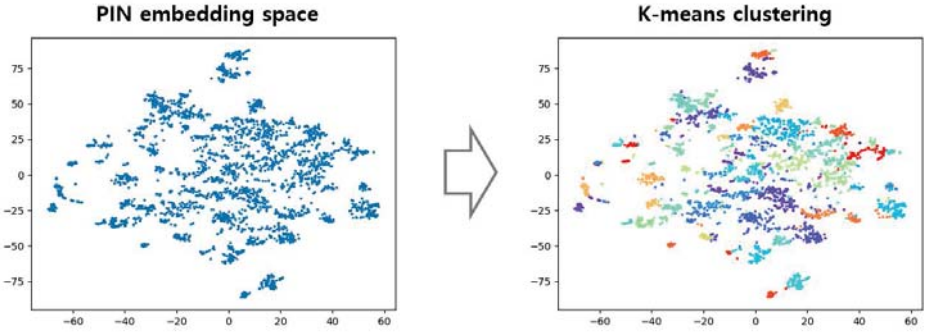

Figure 4.4: An illustration of the subnetwork clustering process.

variables to sample labels. For example, the approach of Lin et al. (Lin *et al.*, 2017) uses an explicit model in which a protein interaction network structure is embedded within a neural network architecture. This approach is useful for improving the performance of the backend prediction model, but it is limited in that the resulting data representation does not provide a biological interpretation. In this study, the implicit model (Conte *et al.*, 2013) was used. In the implicit model, the kernel function is trained to define the distance between

samples, and the kernel's objective function is specified so that the distance between samples can represent the label difference between samples. The training process is as follows (Figure 4.5).

- First, input instances for training the kernel function are collected by a pairwise sample comparison, and each pair of samples is considered an instance.
- For each instance, distances are measured for each cluster. Since each cluster is considered as a vector of the expression values of the included genes, the distance means the distance between vectors.
- Sample label differences (ie, equality: 0 and other: 1) are assigned to each sample pair, which is used as a target variable when training the kernel function.
- The learning algorithm used in the kernel is a non-negative least squares (NNLS) regression, which is implemented in the Python library scipy optimize nnls (Lawson and Hanson, 1995; Bro and De Jong, 1997). The NNLS problem is to find a vector d (K) that minimizes the following expression (4.1) for given Z $(N \times K)$ and x $(N \times 1)$. Here the d is the vector of weights that corresponds to K clusters and Z is the inter-sample distances at each N instances K clusters and x represents the label difference at each sample pair instances.

$$\left\| x - Zd^2 \right\| \tag{4.1}$$

  If d_m is a m-th element of d and $d_m > 0$, then d_m, which is the weight of the m-th cluster, represents the importance of that subnetwork community. Where 0 means that the difference in gene expression in the cluster

does not affect the label difference, and a larger value indicates greater significance.

- After training the kernel, the distance between samples can be measured as a weighted sum of the distance of each cluster of two samples.

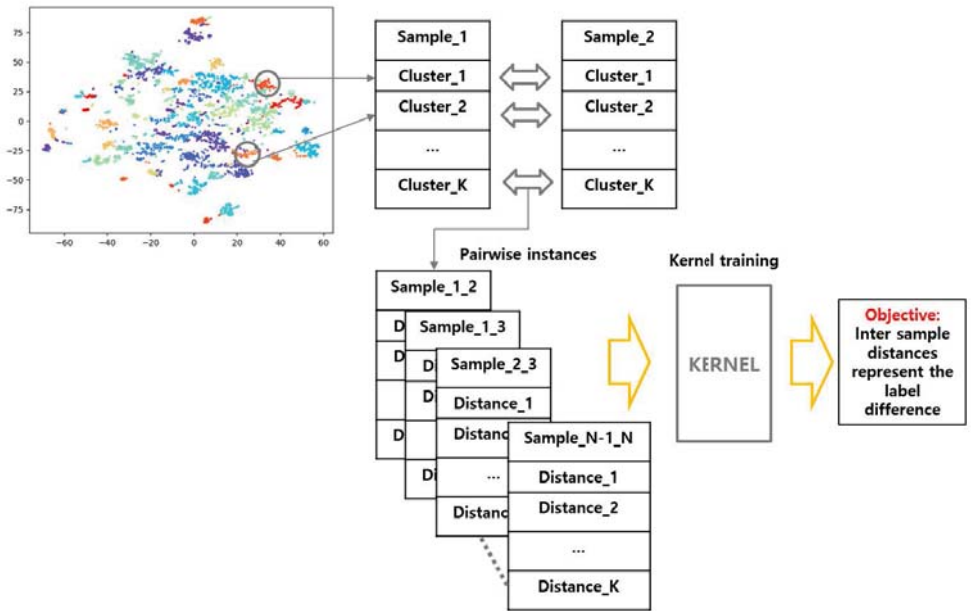

Figure 4.5: An illustration showing the kernel function training process.

Construction of autoencoder for sample embedding calculation

Because the trained kernel functions provide only the distance between samples, an additional step is needed to generate reduced embedding for each RNA-seq sample. An autoencoder model was devised to calculate the sample embedding. The process is as follows (Figure 4.6).

- First, the distance between all training samples is measured using the

trained kernel. Each sample can then be defined as a vector of distances for all other training samples.

- An autoencoder model (Bengio *et al.*, 2009) is created that uses the vectors of distances as input values and has the neural network architecture specified by the user.
- After training the neural network, the embedding of each sample can be calculated by taking the value of the bottleneck layer after forward propagation.
- When data that has not yet been observed is input, the sample is first measured for all training samples and the vector of distances is entered into the autoencoder to generate the embedding.

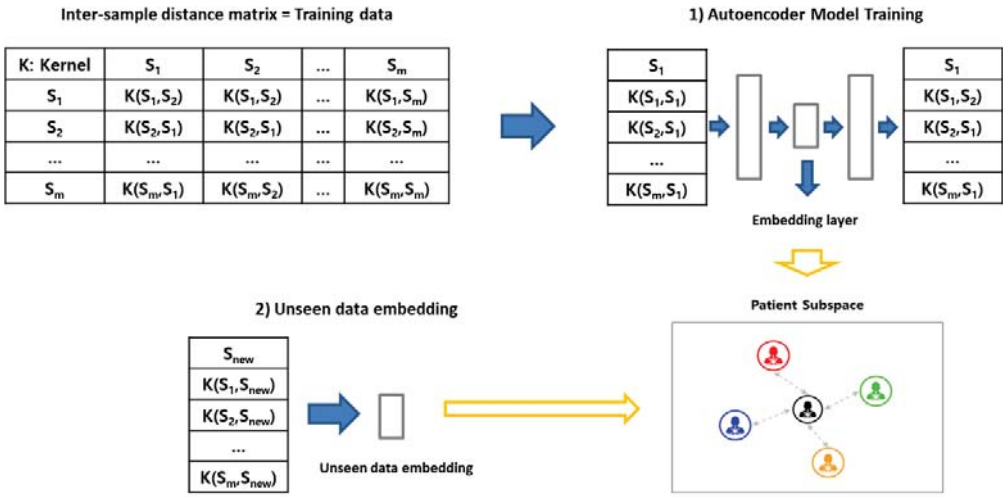

Figure 4.6: An illustration showing the autoencoder training process process.

4.4 Results & Discussion

4.4.1 Lymph node metastasis in early oral cancer

Tumor2Vec was tested in a cancer study to predict lymph node metastasis in early oral cancer. The TCGA-HNSC dataset was used (Network *et al.*, 2015). The data configuration is as follows.

Materials

Of the 566 RNA-seq samples from TCGA-HNSC, only early oral carcinoma samples with lymph node metastasis information were used. That is, tumor samples classified as the oral tongue, alveolar ridge, hard plate, floor of mouth, buccal mucosa, and oral cavity were used. The criterion for the early disease is the pathological T stage within $T1 \sim T2$. In conclusion, there are 60 early oral cancer samples available for lymph node metastasis, 28 of which are positive for lymph node metastasis and 32 negatives. The PIN for subnetwork extraction was extracted from the STRING protein-protein interaction (PPI) network database (Szklarczyk *et al.*, 2014). Interaction edges are filtered with a combined score of 900 to eliminate low confidence interactions.

Results of analysis

The purpose of this analysis is to identify subnetwork level expression patterns (ie, features) that affect lymph node metastasis in early oral cancer. The processing of data is as follows.

- Gene expression profiles of 60 oral cancer samples generated by TCGA were collected, which were measured by RSEM(Li and Dewey, 2011). These cancer samples were normalized using 13 normal oral tissue samples collected from TCGA. In this case, Z-normalization was used.

- Since the available samples are relatively small, only 4,307 genes selected as cancer hallmark gene set in MSigDB were used (Subramanian *et al.*, 2005). Of these, 865 genes are excluded because they are not PPI related to other genes on the STRING PIN (based on edge score >900 cut). As a result, 3,442 genes were used.

- The figure 4.7 is the result of using 91 clusters, and the optimal number of clusters was determined within the range of $10 \sim 200$ by 5-fold cross-validation.

- After the kernel was trained by this configuration, its weight indicates the functional significance of each subnetwork and is displayed as a heatmap, as shown in Figure 4.7-b. Then an autoencoder was created, in which sample embeddings were calculated in two dimensions for visualization and the results are shown in Figure 4.7-a.

- A simple classification model called Nearest Centroid classifier (Tibshirani *et al.*, 2002) was created to test how well the generated two-dimensional sample embeddings distinguish sample labels. The results were 78.3% of the training accuracy and 73.3% of the test accuracy.

Interpretation of subnetwork features

Feature importance of each cluster (ie, subnetwork feature) can be extracted from the trained kernel. The clusters with the top three high scores are listed in the table 4.1. The KEGG_TOP3 column contains the geneset enrichment results from Enrichr (Kuleshov *et al.*, 2016), which lists the top three enrichment score pathways. The genes in each cluster are closely linked according to the STRING PPI (Figure 4.8, 4.9, 4.10). This indicates that graph embedding based

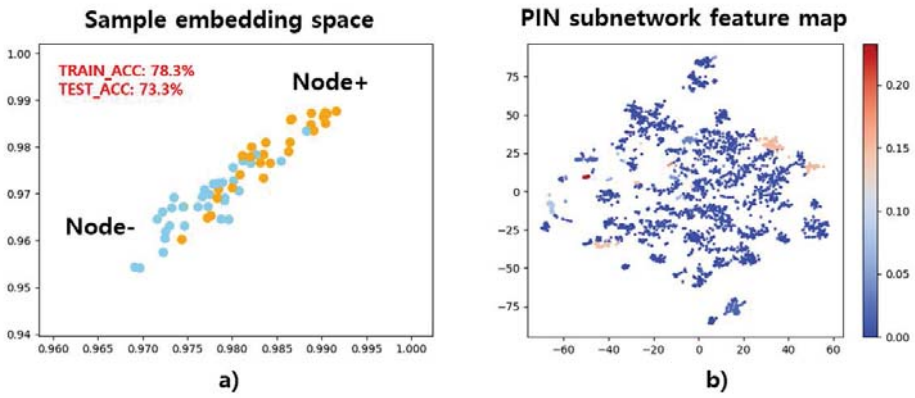

Figure 4.7: Two plots for the results of early oral cancer analysis.

clustering captures the original PIN structure well.

Cluster 1: Subnetwork to regulate leukocyte cell adhesion

Cluster 1 contains genes associated with the tight junction, cell adhesion, and leukocyte migration known to be closely associated with lymph node metastasis in oral cancer (van den Brand *et al.*, 2010; Kudo *et al.*, 2004,?).

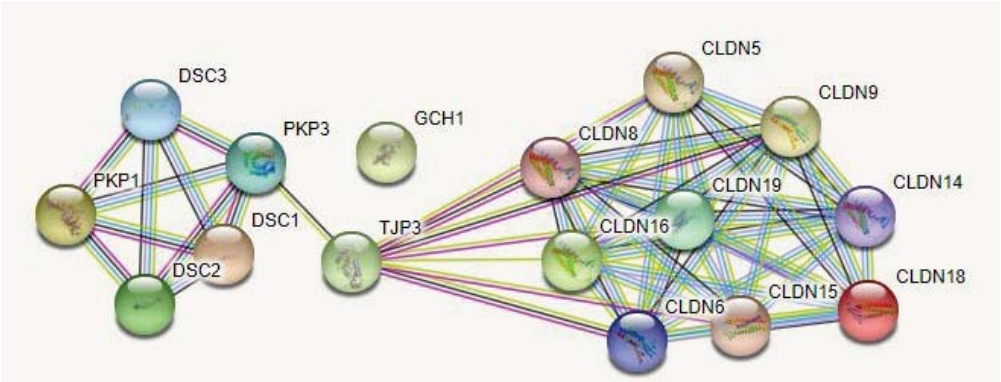

Figure 4.8: A plot showing the STRING PPI interaction between genes in Cluster 1.

Cluster	Kernel Weight	Size	KEGG_TOP3
1	0.233007846	16	Tight junction(10) Cell adhesion molecules (CAMs) (9) Leukocyte transendothelial migration(9)
2	0.185200417	5	Arginine and proline metabolism(3) Glycine, serine and threonine metabolism(3) Tight junction(2)
3	0.15502751	71	Th17 cell differentiation (18) Inflammatory bowel disease (IBD) (16) Cytokine-cytokine receptor interaction (46)

Table 4.1: A table of KEGG enrichment results for Top 3 important subnetwork features.

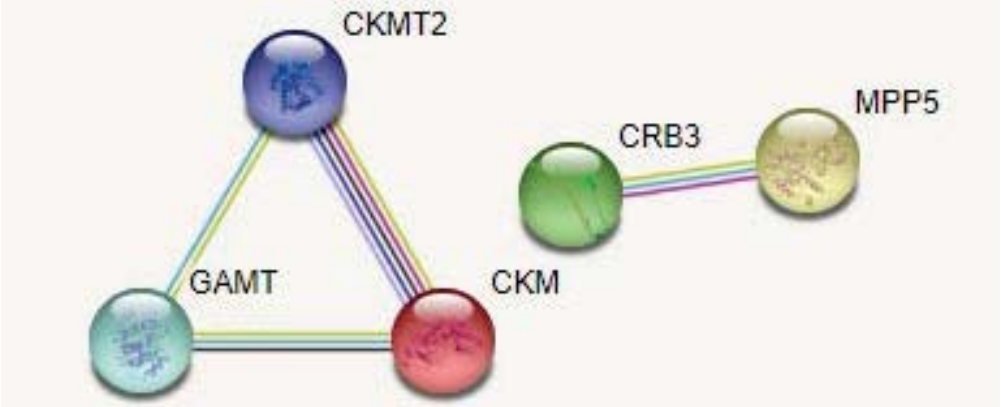

Figure 4.9: A plot showing the STRING PPI interaction between genes in Cluster 2.

4.5 Conclusion

Current dimensional reduction techniques have limitations in that they do not provide a biological interpretation. Tumor2Vec is a machine learning model de-

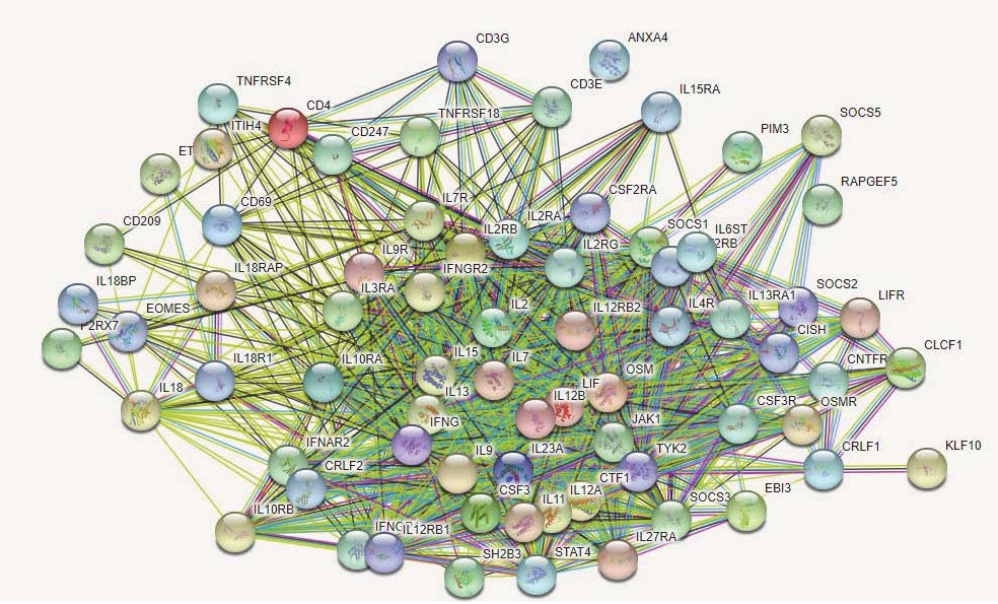

Figure 4.10: A plot showing the STRING PPI interaction between genes in Cluster 3.

veloped to extract subnetwork features that best describe biological phenotype while considering interactions among subnetworks in the training phase. It was tested to identify subnetwork features associated with lymph node metastasis with early oral cancer data. It was able to reproduce clinical knowledge and identify potential subnetwork markers.

Chapter 5

Conclusion

Due to the complex regulatory system, the transcriptome is essentially a mixture containing various transcriptomic variations. This often makes it difficult to see an overall picture of transcriptomic events that regulate biological phenotypes. The goal of my doctoral study was to eliminate the barriers to decoding and utilizing RNA-seq to uncover the landscape of key transcriptomic events. Three key challenges have been addressed using machine learning techniques. Each challenge is summarized as follows:

1. false-positives in RNA editing calls
2. Absence of a model for measuring spliceomic intratumor heterogeneity considering complex cancer spliceome
3. Lack of biological interpretation of dimension reduction techniques using gene expression

In the first study, RDDpred, a condition-specific machine learning model

for filtering false-positive RNA editing calls in RNA-seq data, was developed. There have been limitations to existing methods such as non-machine learning methods lacking generality and machine learning methods requiring extensive proactive experimental validation. RDDpred is a machine learning technique that overcomes these limitations. It uses prior knowledge bases to extract training samples directly from the input data and then generates machine learning predictors specific to the input conditions. RDDpred was tested using the results of two previous studies and showed good results by significantly reducing the false-positive rate while reproducing most of the residues reported in both studies.

In the second study, SpliceHetero, an information-theoretic approach for measuring spliceomic intratumor heterogeneity from bulk tumor RNA-seq data, was developed to solve technical problems caused by complex cancer spliceome. Despite studies that show intercellular differences at the spliceome level, the clinical effect of sITH has not been studied sufficiently because there is no sITH model. SpliceHetero is a sITH model based on local analysis approach that avoids transcriptome assembly which is not easy in cancer RNA-seq. The proposed model was extensively tested for its performance using synthetic data, xenograft tumor data, and TCGA pan-cancer data. As a result, sITH has shown a strong association with cancer progression and clonal heterogeneity as well as clinically relevant features such as cancer progression, survival outcome, and PAM50 subtype.

In the last study, Tumor2Vec, a supervised learning algorithm for extracting subnetwork representations of cancer RNA-seq data using protein interaction networks, was developed. Current dimensional reduction techniques have limitations in that they do not provide a biological interpretation. Tumor2Vec is a machine learning model developed to extract subnetwork features that best de-

scribe biological phenotype while considering interactions among subnetworks in the training phase. It was tested to identify subnetwork features associated with lymph node metastasis with early oral cancer data. It was able to reproduce clinical knowledge and identify potential subnetwork markers.

In conclusion, my doctoral study challenged three major barriers in decoding and utilizing RNA-seq using machine learning techniques. It contributed to the field of bioinformatics by providing solutions to key challenges and opened the way to integrate three transcriptomic domains (ie, RNA editing, alternative splicing, and gene expression) to see an overall picture of transcriptomic events.

Bibliography

Azad, R. K. and Li, J. (2012). Interpreting genomic data via entropic dissection. *Nucleic acids research*, **41**(1), e23–e23.

Bahn, J. H., Lee, J.-H., Li, G., Greer, C., Peng, G., and Xiao, X. (2012). Accurate identification of a-to-i rna editing in human by transcriptome sequencing. *Genome research*, **22**, 142–150.

Bass, B., Hundley, H., Li, J. B., Peng, Z., Pickrell, J., Xiao, X. G., and Yang, L. (2012). The difficult calls in rna editing. *Nature biotechnology*, **30**(12), 1207.

Bengio, Y. *et al.* (2009). Learning deep architectures for ai. *Foundations and trends® in Machine Learning*, **2**(1), 1–127.

Boland, C. R. and Goel, A. (2005). Somatic evolution of cancer cells. In *Seminars in cancer biology*, volume 15, pages 436–450. Elsevier.

Bro, R. and De Jong, S. (1997). A fast non-negativity-constrained least squares algorithm. *Journal of Chemometrics: A Journal of the Chemometrics Society*, **11**(5), 393–401.

Capra, J. A. and Singh, M. (2007). Predicting functionally important residues from sequence conservation. *Bioinformatics*, **23**(15), 1875–1882.

Carter, S. L., Cibulskis, K., Helman, E., McKenna, A., Shen, H., Zack, T., Laird, P. W., Onofrio, R. C., Winckler, W., Weir, B. A., *et al.* (2012). Absolute quantification of somatic dna alterations in human cancer. *Nature biotechnology*, **30**(5), 413.

Chen, H., Lin, F., Xing, K., and He, X. (2015). The reverse evolution from multicellularity to unicellularity during carcinogenesis. *Nature communications*, **6**, 6367.

Chiu, Y.-L., Soros, V. B., Kreisberg, J. F., Stopak, K., Yonemoto, W., and Greene, W. C. (2010). Cellular apobec3g restricts hiv-1 infection in resting cd4+ t cells. *Nature*, **466**, 276–276.

Cichocki, A. and Phan, A.-H. (2009). Fast local algorithms for large scale nonnegative matrix and tensor factorizations. *IEICE transactions on fundamentals of electronics, communications and computer sciences*, **92**(3), 708–721.

Climente-González, H., Porta-Pardo, E., Godzik, A., and Eyras, E. (2017). The functional impact of alternative splicing in cancer. *Cell reports*, **20**(9), 2215–2226.

Conte, D., Ramel, J.-Y., Sidère, N., Luqman, M. M., Gaüzère, B., Gibert, J., Brun, L., and Vento, M. (2013). A comparison of explicit and implicit graph embedding methods for pattern recognition. In *International Workshop on Graph-Based Representations in Pattern Recognition*, pages 81–90. Springer.

Degner, J. F., Marioni, J. C., Pai, A. A., Pickrell, J. K., Nkadori, E., Gilad, Y.,

and Pritchard, J. K. (2009). Effect of read-mapping biases on detecting allele-specific expression from rna-sequencing data. *Bioinformatics*, **25**, 3207–3212.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. (2013). Star: ultrafast universal rna-seq aligner. *Bioinformatics*, **29**, 15–21.

Dvinge, H. and Bradley, R. K. (2015). Widespread intron retention diversifies most cancer transcriptomes. *Genome medicine*, **7**(1), 45.

Engström, P. G., Steijger, T., Sipos, B., Grant, G. R., Kahles, A., Alioto, T., Behr, J., Bertone, P., Bohnert, R., Campagna, D., *et al.* (2013). Systematic evaluation of spliced alignment programs for rna-seq data. *Nature methods*, **10**(12), 1185.

Eswaran, J., Horvath, A., Godbole, S., Reddy, S. D., Mudvari, P., Ohshiro, K., Cyanam, D., Nair, S., Fuqua, S. A., Polyak, K., *et al.* (2013). Rna sequencing of cancer reveals novel splicing alterations. *Scientific reports*, **3**, 1689.

Fan, P., Lin, Q.-H., Guo, Y., Zhao, L.-L., Ning, H., Liu, M.-Y., and Wei, D.-Q. (2018). The ppi network analysis of mrna expression profile of uterus from primary dysmenorrheal rats. *Scientific reports*, **8**(1), 351.

Galeano, F., Rossetti, C., Tomaselli, S., Cifaldi, L., Lezzerini, M., Pezzullo, M., Boldrini, R., Massimi, L., Di Rocco, C., Locatelli, F., *et al.* (2013). Adar2-editing activity inhibits glioblastoma growth through the modulation of the cdc14b/skp2/p21/p27 axis. *Oncogene*, **32**, 998–1009.

Graf, J. F. and Zavodszky, M. I. (2017). Characterizing the heterogeneity of tumor tissues from spatially resolved molecular measures. *PloS one*, **12**(11), e0188878.

Greaves, M. and Maley, C. C. (2012). Clonal evolution in cancer. *Nature*, **481**(7381), 306.

Haas, B. J. and Zody, M. C. (2010). Advancing rna-seq analysis. *Nature biotechnology*, **28**(5), 421.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, **11**, 10–18.

Heap, G. A., Yang, J. H., Downes, K., Healy, B. C., Hunt, K. A., Bockett, N., Franke, L., Dubois, P. C., Mein, C. A., Dobson, R. J., *et al.* (2010). Genome-wide analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome resequencing. *Human molecular genetics*, **19**, 122–134.

Hoffman, M., Bach, F. R., and Blei, D. M. (2010). Online learning for latent dirichlet allocation. In *advances in neural information processing systems*, pages 856–864.

Jardim-Perassi, B. V., Alexandre, P. A., Sonehara, N. M., de Paula-Junior, R., Júnior, O. R., Fukumasu, H., Chammas, R., Coutinho, L. L., and de Campos Zuccari, D. A. P. (2019). Rna-seq transcriptome analysis shows antitumor actions of melatonin in a breast cancer xenograft model. *Scientific reports*, **9**(1), 966.

Jayasinghe, R. G., Cao, S., Gao, Q., Wendl, M. C., Vo, N. S., Reynolds, S. M., Zhao, Y., Climente-González, H., Chai, S., Wang, F., *et al.* (2018). Systematic analysis of splice-site-creating mutations in cancer. *Cell reports*, **23**(1), 270–281.

Joyce, J. M. (2011). Kullback-leibler divergence. In *International Encyclopedia of Statistical Science*, pages 720–722. Springer.

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2016). Kegg: new perspectives on genomes, pathways, diseases and drugs. *Nucleic acids research*, **45**(D1), D353–D361.

Kiran, A. and Baranov, P. V. (2010). Darned: a database of rna editing in humans. *Bioinformatics*, **26**, 1772–1776.

Kruskal, J. B. (1964). Nonmetric multidimensional scaling: a numerical method. *Psychometrika*, **29**(2), 115–129.

Kudo, Y., Kitajima, S., Ogawa, I., Hiraoka, M., Sargolzaei, S., Keikhaee, M. R., Sato, S., Miyauchi, M., and Takata, T. (2004). Invasion and metastasis of oral cancer cells require methylation of e-cadherin and/or degradation of membranous β-catenin. *Clinical Cancer Research*, **10**(16), 5455–5463.

Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S. L., Jagodnik, K. M., Lachmann, A., *et al.* (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic acids research*, **44**(W1), W90–W97.

Kumari, K., Keshari, S., Sengupta, D., Sabat, S. C., and Mishra, S. K. (2017). Transcriptome analysis of genes associated with breast cancer cell motility in response to artemisinin treatment. *BMC cancer*, **17**(1), 858.

Lawson, C. L. and Hanson, R. J. (1995). *Solving least squares problems*, volume 15. Siam.

Li, B. and Dewey, C. N. (2011). Rsem: accurate transcript quantification from

rna-seq data with or without a reference genome. *BMC bioinformatics*, **12**(1), 323.

Li, H. (2011). A statistical framework for snp calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, **27**, 2987–2993.

Li, J. B., Levanon, E. Y., Yoon, J.-K., Aach, J., Xie, B., LeProust, E., Zhang, K., Gao, Y., and Church, G. M. (2009). Genome-wide identification of human rna editing sites by parallel dna capturing and sequencing. *Science*, **324**(5931), 1210–1213.

Licht, K. and Jantsch, M. F. (2016). Rapid and dynamic transcriptome regulation by rna editing and rna modifications. *J Cell Biol*, **213**(1), 15–22.

Lin, C., Jain, S., Kim, H., and Bar-Joseph, Z. (2017). Using neural networks for reducing the dimensions of single-cell rna-seq data. *Nucleic acids research*, **45**(17), e156–e156.

Lin, J. (1991). Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, **37**(1), 145–151.

Lin, K.-H., Huang, M.-Y., Cheng, W.-C., Wang, S.-C., Fang, S.-H., Tu, H.-P., Su, C.-C., Hung, Y.-L., Liu, P.-L., Chen, C.-S., *et al.* (2018). Rna-seq transcriptome analysis of breast cancer cell lines under shikonin treatment. *Scientific reports*, **8**(1), 2672.

Liu, Y., Gonzàlez-Porta, M., Santos, S., Brazma, A., Marioni, J. C., Aebersold, R., Venkitaraman, A. R., and Wickramasinghe, V. O. (2017). Impact of alternative splicing on the human proteome. *Cell reports*, **20**(5), 1229–1241.

Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, **9**(Nov), 2579–2605.

Marusyk, A. and Polyak, K. (2010). Tumor heterogeneity: causes and consequences. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, **1805**(1), 105–117.

Mazor, T., Pankov, A., Song, J. S., and Costello, J. F. (2016). Intratumoral heterogeneity of the epigenome. *Cancer cell*, **29**(4), 440–451.

McGettigan, P. A. (2013). Transcriptomics in the rna-seq era. *Current opinion in chemical biology*, **17**(1), 4–11.

McGranahan, N. and Swanton, C. (2017). Clonal heterogeneity and tumor evolution: past, present, and the future. *Cell*, **168**(4), 613–628.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., *et al.* (2010). The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome research*, **20**, 1297–1303.

Minka, T. P. (2001). Automatic choice of dimensionality for pca. In *Advances in neural information processing systems*, pages 598–604.

MIT (2011). Wgsim.

Mo, F., Wyatt, A. W., Sun, Y., Brahmbhatt, S., McConeghy, B. J., Wu, C., Wang, Y., Gleave, M. E., Volik, S. V., and Collins, C. C. (2014). Systematic identification and characterization of rna editing in prostate tumors. *PloS one*, **9**(7), e101431.

Morris, L. G., Riaz, N., Desrichard, A., Şenbabaoğlu, Y., Hakimi, A. A., Makarov, V., Reis-Filho, J. S., and Chan, T. A. (2016). Pan-cancer analysis

of intratumor heterogeneity as a prognostic determinant of survival. *Oncotarget*, **7**(9), 10051.

Network, C. G. A. *et al.* (2012). Comprehensive molecular portraits of human breast tumours. *Nature*, **490**(7418), 61.

Network, C. G. A. *et al.* (2015). Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature*, **517**(7536), 576.

Ng, A. Y., Jordan, M. I., and Weiss, Y. (2002). On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems*, pages 849–856.

Nowell, P. C. (1976). The clonal evolution of tumor cell populations. *Science*, **194**(4260), 23–28.

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., *et al.* (2015). Reference sequence (refseq) database at ncbi: current status, taxonomic expansion, and functional annotation. *Nucleic acids research*, **44**(D1), D733–D745.

Park, Y., Lim, S., Nam, J.-W., and Kim, S. (2016). Measuring intratumor heterogeneity by network entropy using rna-seq data. *Scientific reports*, **6**, 37767.

Parker, J. S., Mullins, M., Cheang, M. C., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., *et al.* (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of clinical oncology*, **27**(8), 1160.

Patel, A. P., Tirosh, I., Trombetta, J. J., Shalek, A. K., Gillespie, S. M., Wakimoto, H., Cahill, D. P., Nahed, B. V., Curry, W. T., Martuza, R. L., *et al.* (2014). Single-cell rna-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, **344**(6190), 1396–1401.

Peng, Z., Cheng, Y., Tan, B. C.-M., Kang, L., Tian, Z., Zhu, Y., Zhang, W., Liang, Y., Hu, X., Tan, X., *et al.* (2012). Comprehensive analysis of rna-seq data reveals extensive rna editing in a human transcriptome. *Nature biotechnology*, **30**, 253–260.

Perozzi, B., Al-Rfou, R., and Skiena, S. (2014). Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 701–710. ACM.

Prahallad, A. and Bernards, R. (2016). Opportunities and challenges provided by crosstalk between signalling pathways in cancer. *Oncogene*, **35**(9), 1073.

Rajan, P., Elliott, D. J., Robson, C. N., and Leung, H. Y. (2009). Alternative splicing and biological heterogeneity in prostate cancer. *Nature Reviews Urology*, **6**(8), 454.

Ramaswami, G. and Li, J. B. (2013). Radar: a rigorously annotated database of a-to-i rna editing. *Nucleic acids research*, page gkt996.

Roth, A., Khattra, J., Yap, D., Wan, A., Laks, E., Biele, J., Ha, G., Aparicio, S., Bouchard-Côté, A., and Shah, S. P. (2014). Pyclone: statistical inference of clonal population structure in cancer. *Nature methods*, **11**(4), 396.

Roweis, S. T. and Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *science*, **290**(5500), 2323–2326.

Sebestyén, E., Singh, B., Miñana, B., Pagès, A., Mateo, F., Pujana, M. A., Valcárcel, J., and Eyras, E. (2016). Large-scale analysis of genome and transcriptome alterations in multiple tumors unveils novel cancer-relevant splicing networks. *Genome research*, **26**(6), 732–744.

Shalek, A. K., Satija, R., Adiconis, X., Gertner, R. S., Gaublomme, J. T., Raychowdhury, R., Schwartz, S., Yosef, N., Malboeuf, C., Lu, D., *et al.* (2013). Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature*, **498**(7453), 236.

Shen, D., Shen, H., Zhu, H., and Marron, J. (2016). The statistics and mathematics of high dimension low sample size asymptotics. *Statistica Sinica*, **26**(4), 1747.

St Laurent, G., Tackett, M. R., Nechkin, S., Shtokalo, D., Antonets, D., Savva, Y. A., Maloney, R., Kapranov, P., Lawrence, C. E., and Reenan, R. A. (2013). Genome-wide analysis of a-to-i rna editing by single-molecule sequencing in drosophila. *Nature structural & molecular biology*, **20**, 1333–1339.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., *et al.* (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, **102**(43), 15545–15550.

Sun, X.-x. and Yu, Q. (2015). Intra-tumor heterogeneity of cancer cells and its implications for cancer treatment. *Acta Pharmacologica Sinica*, **36**(10), 1219.

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K. P., *et al.* (2014).

String v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic acids research*, **43**(D1), D447–D452.

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N. T., Morris, J. H., Bork, P., *et al.* (2018). String v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic acids research*, **47**(D1), D607–D613.

Tenenbaum, J. B., De Silva, V., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *science*, **290**(5500), 2319–2323.

Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, **99**(10), 6567–6572.

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., Van Baren, M. J., Salzberg, S. L., Wold, B. J., and Pachter, L. (2010). Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, **28**(5), 511.

Treutlein, B., Brownfield, D. G., Wu, A. R., Neff, N. F., Mantalas, G. L., Espinoza, F. H., Desai, T. J., Krasnow, M. A., and Quake, S. R. (2014). Reconstructing lineage hierarchies of the distal lung epithelium using single-cell rna-seq. *Nature*, **509**(7500), 371.

Tsai, Y. S., Dominguez, D., Gomez, S. M., and Wang, Z. (2015). Transcriptome-wide identification and study of cancer-specific splicing events across multiple tumors. *Oncotarget*, **6**(9), 6825.

van den Brand, M., Takes, R. P., Blokpoel-deRuyter, M., Slootweg, P. J., and van Kempen, L. C. (2010). Activated leukocyte cell adhesion molecule expression predicts lymph node metastasis in oral squamous cell carcinoma. *Oral oncology*, **46**(5), 393–398.

Wan, Y. and Larson, D. R. (2018). Splicing heterogeneity: separating signal from noise. *Genome biology*, **19**(1), 86.

Wang, D. and Gu, J. (2018). Vasc: Dimension reduction and visualization of single-cell rna-seq data by deep variational autoencoder. *Genomics, proteomics & bioinformatics*, **16**(5), 320–331.

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J. M., Network, C. G. A. R., *et al.* (2013). The cancer genome atlas pan-cancer analysis project. *Nature genetics*, **45**(10), 1113.

Xiang, Y., Ye, Y., Zhang, Z., and Han, L. (2018). Maximizing the utility of cancer transcriptomic data. *Trends in cancer*.

Xiong, Y., You, W., Wang, R., Peng, L., and Fu, Z. (2017). Prediction and validation of hub genes associated with colorectal cancer by integrating ppi network and gene expression data. *BioMed research international*, **2017**.

Yu, W., He, L.-R., Zhao, Y.-C., Chan, M.-H., Zhang, M., and He, M. (2013). Dynamic protein-protein interaction subnetworks of lung cancer in cases with smoking history. *Chinese journal of cancer*, **32**(2), 84.

Yuan, X., Chen, J., Lin, Y., Li, Y., Xu, L., Chen, L., Hua, H., and Shen, B. (2017). Network biomarkers constructed from gene expression and protein-

protein interaction data for accurate prediction of leukemia. *Journal of Cancer*, **8**(2), 278.

Zhang, Q. and Xiao, X. (2015). Genome sequence-independent identification of rna editing sites. *Nature methods*, **12**, 347–350.

Zhu, D., Zhaozu, X., Cui, G., Chang, S., See, Y. X., Lim, M. G. L., Guo, D., Chen, X., Robson, P., Luo, Y., *et al.* (2018). Single-cell transcriptome analysis reveals estrogen signaling augments the mitochondrial folate pathway to coordinately fuel purine and polyamine synthesis in breast cancer cells. *bioRxiv*, page 246363.

초록

진핵 세포 시스템에서는 mRNA 분자가 전사된 이후 완전히 처리되어 단백질로 번역될 때까지 여러 단계의 전사 후 조절 과정을 거치게 된다. 전사 후 조절 과정은 RNA 편집, 선택적 접합, 선택적 아데닐화 등을 포함한다. 즉 어느 한 시점에서 전사체를 들여다보면 그 내부는 다양한 중간체들의 혼합물로 구성되어 있는 것이다. 이러한 복잡한 조절 시스템 때문에 전사체를 전체적인 수준에서 이해하기가 쉽지 않다. 본 학위 연구는 RNA 시퀀싱 데이터를 해독하고 활용하기 위한 기계학습 기법들에 대한 연구이며 RNA 편집, 선택적 접합 및 유전자 발현의 관점에서 수행된 세 가지 연구로 구성된다.

RNA 편집은 ADAR(A=>I) 과 APOBEC(C=>U) 두 가지 효소에 의해 촉매되는 전사 후 RNA 서열 조절 기작이다. RNA 편집은 단백질 활성도, 선택적 접합 및 miRNA 표적 조절 등 다양한 세포 기작을 제어하는 것으로 알려진 중요한 새포 내 조절 시스템이다. RNA 시퀀싱을 이용해 RNA 편집 현상을 검출하는 것은 RNA 편집 현상의 생물학적 기능을 이해하는 데에 매우 중요하다. 문제는 이 과정에서 상당한 양의 위양성이 발생한다는 점이다. 샘플당 수만 개 이상 발생하는 RNA 편집 잔기들 모두를 실험적으로 검증할 수 없기 때문에 이를 걸러내기 위한 전산학적 모델이 요구된다. RDDpred는 RNA 시퀀싱 데이터로부터 RNA 편집 현상을 검출하는 과정에서 발생하는 위양성 잔기들을 기계학습 기술에 기반하여 구분하는 모델이다. RDDpred는 두 개의 기 발표된 RNA 편집 연구 데이터를 이용하여 검증되었다.

RNA 시퀀싱 기술이 활용될 수 있는 또 하나의 복잡한 문제로 접합체 차원에서의 종양 이질성 (ITH) 측정 문제가 있다. ITH는 암 조직을 구성하는 세포 집단의 다양성의 지표이며, 최근 출판된 연구들의 결과는 유전자 발현량 데이터에 기반하여 측정된 전사체 수준에서의 ITH가 암 환자의 예후예측에 유용함을 시사한다.

접합체는 유전자 발현량과 함께 전사체를 구성하는 주요 요소 중 하나이며 따라서 접합체 수준에서 ITH를 측정하는 것은 보다 전체적인 수준에서 전사체 ITH를 연구하기 위한 자연스러운 흐름이다. RNA 시퀀싱 데이터를 이용하여 암 접합체 수준에서 ITH를 측정하는 과정에는 복잡한 접합 패턴과 광범위한 인트론 연장 변이 및 짧은 시퀀싱 판독 길이 등의 심각한 기술적 난관들이 있다. SpliceHetero는 이러한 문제들을 고려하여 접합체 수준에서의 ITH (즉, sITH)를 측정하기 위한 도구이며 내부적으로 정보이론을 활용한다. SpliceHetero는 시뮬레이션 데이터, 이종이식 종양 데이터 및 TCGA pan-cancer 데이터 등을 활용하여 광범위하게 검증되었으며 ITH를 잘 반영하는 것으로 확인되었다. 이뿐 아니라 sITH는 암의 진행과 암 환자의 예후 및 PAM50와 같은 잘 알려진 분자 아형들과도 높은 상관관계를 가지는 것으로 확인되었다.

마지막 연구 주제는 유전자 발현량 데이터에 기반하여 특정 암 표현형에 특이적인 환자 부분 공간을 정의하는 기계학습 알고리즘을 개발하는 것이다. RNA 시퀀싱 데이터는 암 환자의 유전자 발현량 프로파일을 얻는 데에 유용한 도구이지만, 2만 개 이상의 차원을 가진 매우 고차원의 데이터이기 때문에 실질적인 용도로 사용되기 위해서는 그 차원의 크기를 축소할 필요가 있다. 이때 각 유전자들은 복잡하지만 고유한 방식으로 서로 상호작용한다는 점을 이용할 수 있다. 실험적으로 검증된 단백질 간의 상호작용 정보를 모아 네트워크 형태로 묶은 것을 단백질 상호작용 네트워크 (혹은 PIN)라 부른다. 이 PIN을 활용하여 RNA 시퀀싱 데이터의 차원을 줄이면서도 데이터로부터 생물학적으로 유의미한 특징들을 추출할 수 있다. Tumor2Vec은 이렇게 추출된 PIN 수준의 특징들을 활용하여 특정 암 표현형에 특이적인 환자 부분 공간을 정의한다. Tumor2Vec은 조기 구강 암에서 림프절 전이를 예측하기 위한 파일럿 연구에 적용되었으며 그 결과 RNA 시퀀싱 데이터의 차원을 줄여 림프절 전이 예측 모델을 생성했고 이 과정에서 암 표현형을 잘 설명하는 PIN 수준의 특징들을 보존하는 데에도 성공했다.

주요어: RNA 시퀀싱, RNA 편집, 선택적 접합, 유전자 발현, 기계학습, 정보이론, 그래프 임베딩, 차원 축소, 오토인코더

학번: 2013-23006